

SPEED AND PERFORMANCE DIFFERENCES AMONG COMPUTER-BASED AND PAPER-PENCIL TESTS

SHAWN M. BODMANN

University of Wisconsin

DANIEL H. ROBINSON

University of Texas

ABSTRACT

This study investigated the effect of several different modes of test administration on scores and completion times. In Experiment 1, paper-based assessment was compared to computer-based assessment. Undergraduates completed the computer-based assessment faster than the paper-based assessment, with no difference in scores. Experiment 2 assessed three different computer interfaces that provided students various levels of flexibility to change and review answers. No difference in scores was observed among the three modes, but students completed the least-flexible mode faster than the other two modes. It appears that less flexible test modes are faster and do not result in poorer performance than more flexible modes.

INTRODUCTION

As computers become increasingly available in educational settings, it is likely that teachers will use them to administer tests (Trotter, 2001). Computer-based tests (CBTs) offer several advantages over traditional paper-and-pencil or paper-based tests (PPTs), even if the computer version is a simple, non-adaptive replication of the paper version. Once set up, CBTs are easier to administer than PPTs. CBTs offer the possibility of instant grading (Bugbee & Bernt, 1990; Inouye & Bunderson, 1986), and, if part of a larger administrative system, automatic tracking and averaging of grades. CBTs are easier to manipulate in order to reduce cheating. Testing conditions can be standardized, and the sequence of items is

easily manipulated (Inouye & Bunderson, 1986). In some cases, CBTs can be administered in a way that allows students to choose when they take a test (Bugbee & Bernt, 1990). CBTs are also capable of collecting metrics that PPTs cannot, such as test and item latency rates (Inouye & Bunderson, 1986; Olsen, Maynes, Slawson, & Ho, 1989).

Computerized adaptive tests (CATs) offer several advantages over non-adaptive CBTs, but their development involves myriad practical challenges (Mills & Stocking, 1996). Inouye and Bunderson (1986) estimated that 90% of computerized testing is in non-adaptive CBT formats. Despite the age of this estimate, overcoming the challenges to implementing CATs in a public school setting, where 82% of teachers report lack of sufficient release time to learn how to use computers (U.S. Department of Education, 2000), seems unlikely. Therefore, we suspect that the majority of CBTs used in the classroom are non-adaptive.

As more tests are converted to computer administration, the question of whether the mode of administration affects scores becomes increasingly important to assess. Accurate measurement is always more desirable than inaccurate measurement, for obvious ethical, administrative, and scientific reasons, so reducing test mode effects is beneficial to schools and students alike. A very large body of research into this question already exists, but unfortunately, this research has produced mixed results. Some studies have found lower scores on CBTs compared with PPTs (Frederico, 1989; Lee, Moreno, & Sympson, 1986; Mazzeo, Druesne, Raffield, Checketts, & Muelstein, 1991; Russell, 1999), higher scores on CBTs compared with PPTs (Bocij & Greasley, 1999; Bugbee & Bernt, 1990; Clariana & Wallace, 2002; DeAngelis, 2000; Pomplun, Frey, & Becker, 2002), or no test mode effects at all (Mason, Patry, & Bernstein, 2001; Mead & Drasgow, 1993; Neuman, & Baydoun, 1998; Olsen, Maynes, Slawson, & Ho, 1989; Ward, Hooper, & Hannafin, 1989; Weinberg, 2001).

Researchers have advanced various hypotheses to explain such test mode effects. Lee, Moreno, and Sympson (1986) cite time available for testing, test difficulty, cognitive processes required by the test, and the absence or presence of a test administrator as factors in test mode effects. Russell (1999) found that students with lower-than-average keyboarding skills performed worse on CBTs. Clariana and Wallace (2002) found gender, competitiveness, and computer familiarity were unrelated to test mode effects, but content familiarity was.

Perhaps the best-supported reason for test mode effects is the difference in flexibility of test modes. For example, some CBTs do not provide the same level of flexibility as PPTs. Specifically, some computer interfaces do not allow the student to skip, review, and/or change answers. Spray, Ackerman, Reckase, and Carlson (1989) claimed that differences in flexibility between CBTs and PPTs are responsible for test mode effects. Mason, Patry, and Bernstein (2001) also found evidence supporting a test mode effect caused by differences in flexibility.

Empirical evidence exists to support the hypothesis that differences in flexibility would produce differences in test scores. There have been several studies on

the effect of changing answers on PPTs, and the clear consensus is that changing answers on multiple-choice tests slightly increases scores (Kruger, Wirtz, & Miller, in review; Mueller & Wasser, 1977; Schwarz, McMorris, & DeMers, 1991; Vispoel, 1998). Therefore, it would follow that CBTs that do not allow the student to change answers may result in lower test scores than the same tests administered in a paper format.

Most of the research investigating the equivalence between CBTs and PPTs focuses on scores, but also of interest is whether students complete CBTs faster than PPTs. Greaud and Green (1986) and Olson, Maynes, Slawson, and Ho (1989) found CBTs were faster, but Zandvliet and Farragher (1997) found that CBTs took longer for people with minimal computer skills. Dimock (1991) found that CBTs took longer the first time students used them, but not on subsequent administrations. Bugbee and Bernt (1990) found that slightly more students who took a CBT desired more than the allotted time to complete it than students who took a PPT. Participants in the Dimock (1991), Greaud and Green (1986), and Zandvliet and Farragher (1997) studies were not familiar with the testing interface before the study was conducted. Students in the Bugbee and Bernt (1999) study were assumed to be computer literate, but it is unclear whether they had experience with the specific testing interface used in that study. It is unclear whether students in the Olson, Maynes, Slawson, and Ho (1989) study were computer-literate or had prior experience with the testing interface.

EXPERIMENT 1

Experiment 1 was designed to detect differences in scores and completion times between a PPT and a CBT for computer-literate students with prior experience with the CBT interface.

Method

Participants and Design

Fifty-five students enrolled in an undergraduate educational psychology course participated. Every student in the class had previously taken three tests on computer, so they were familiar with the interface used in this study. Students were randomly assigned to one of two experimental groups.

Materials

All tests covered course content and contributed to the students' overall course grades. Both CBTs and PPTs contained 30 items, with four answer choices per question and a 35-minute time limit. We used the CBT system available through ActiveInk, a Web-based course management system. This system displays a single question at a time, including the stem and answer choices. Students select an

answer via standard radio buttons, and then click a second button to submit the answer. Radio button selections can be changed multiple times, but once an answer is submitted, it cannot be reviewed or changed. Questions cannot be skipped or returned to later. The CBT maintains and displays a clock showing the students' total elapsed testing time. The PPTs contained the exact same questions as the corresponding CBTs, in exactly the same order. To make the PPTs similar to typical ones used in classrooms, they displayed several questions (about six) per page.

Procedure

Approximately half the class (28 students) took the first test on the computer, whereas the rest (27 students) took the first test on paper. The procedures were switched for the second test, with the first group receiving PPTs and the second group receiving CBTs. The two tests were administered two weeks apart and covered different material. Students took the CBTs in a computer lab that was different than the classroom where they normally attended class. To facilitate timing the PPT administration, tests were distributed and students were instructed not to start the test until a signal was given. The elapsed time for each student was recorded as individual tests were handed in (10 seconds were subtracted from each time to allow for the time to walk to the proctor's table, after averaging the times during a simulated testing session). The PPTs were administered in the same room as the students normally attended lectures. To control for any contextual retrieval advantages of the PPT group, students were seated facing the opposite side of the room they faced during normal lectures.

Results and Discussion

Dependent *t*-tests were conducted for both the test scores and test times. Table 1 displays the mean scores and times for each test mode. For the test scores, there was no difference between the PPT and CBT, $t(53) = .04, p = .97$. However, for the test times, the PPT took an average of almost four minutes longer than the CBT, $t(53) = 3.95, p < .001$.

Experiment 1 found no test mode effect for students' scores. However, a test mode effect for completion time was found, with students completing the CBTs in less time than the PPTs. One reviewer suggested that multiple items per page on the PPT might have caused cognitive dissonance, which then could have caused increased completion times. We find this explanation to be unlikely. First, it is unclear how multiple questions covering related course content would act as dissonant cognitions. Second, multiple items on the same page seems by far the norm for PPT tests, and participants were assumed to have much experience with such formats. We thought it much more likely that the added flexibility of the PPT caused the increased completion times. We designed Experiment 2 to help address this issue.

Table 1. Mean Test Scores (Out of 30) and Test Times (in Minutes) for Both Test Modes in Experiment 1

	Test scores		Test times	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
PPT	23.69	3.47	27.2	6.34
CBT	23.70	3.51	23.4	5.92
Probability	.97		<.001	

EXPERIMENT 2

Experiment 2 was designed to eliminate some of the variability between the test modes used in Experiment 1. We wanted to focus on any score or time differences caused by the ability to skip, review, and change answers. To do so, we eliminated the PPT part of Experiment 1 and, instead, used three CBTs with different levels of flexibility.

Method

Participants and Design

Fifty-eight students in a different section of the same educational psychology undergraduate course participated. It is unknown whether students had prior experience with the test administration software. Students were randomly assigned to one of three experimental groups.

Materials

As in Experiment 1, all tests were 30 items long, with four answer choices per question and a 35-minute time limit. We used three different computer interfaces that were available as part of the WebCT course management system:

Interface #1 very closely approximates PPTs. This interface presents the student with all 30 questions at once, and allows navigation through the test items via a scroll bar. Just like PPTs, this interface allows infinite skipping, review, and change capabilities. This interface will be referred to as the “all-at-once, scroll” condition.

Interface #2 displays a single question and answer set at a time (on the screen), but allows the student to review and change answers before submitting the entire test. This interface provides the same capabilities as PPTs, though through a slightly different interface (a single question at a time rather than

several questions on a printed page). This interface will be referred to as the “one-at-a-time, revisit” condition.

Interface #3 is very similar to the computer interface used in Experiment 1. Students are presented with a single question at a time. The student selects one of the answers, commits the answer, and moves to the next question. No review or changing of answers is allowed. This interface is referred to as the “one-at-a-time, no revisit” condition.

Procedure

A total of three different tests were administered over a period of six weeks. Each group of students received one of the three interface conditions for each test. The order of the test modes was counterbalanced such that one-third of the students received Interface #1 first, one-third received Interface #2 first, and one-third received Interface #3 first.

Results and Discussion

Repeated measures ANOVAs were conducted on both the scores and times for the three test modes. Table 2 displays mean scores and times for the three modes. Similar to Experiment 1, there were no differences in test scores, $F(2, 114) = 0.77$, $MSE = 6.76$, $p = .46$. However, again similar to Experiment 1, there was a difference in test times, $F(2, 114) = 12.24$, $MSE = 1346.5$, $p < .001$. Post hoc Fisher LSD tests revealed that the one-at-a-time, no revisit version was, on average, about 2.5 to four minutes faster than the other two versions.

Consistent with Experiment 1, we failed to find any test mode effect on test scores but did find a test mode effect on test times. As expected, differences in flexibility affected test completion time. The one-at-a-time, no revisit

Table 2. Mean Test Scores (Out of 30) and Test Times (in Minutes) for the Three Test Modes in Experiment 2

	Test scores		Test times	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
All-at-once, scroll	22.07	3.50	26.6	8.44
One-at-a-time, revisit	22.52	3.33	24.9	5.24
One-at-a-time, no revisit	21.95	3.54	22.3	4.67
<i>F</i> -ratio (2, 114 <i>df</i>)	0.77		12.24	
Probability	.46		<.001	

condition was the least flexible of the three conditions, and students completed it the fastest. This indicates that the increased flexibility of the other conditions did cause some of the increase in completion time, but did not have an effect on student scores.

GENERAL DISCUSSION

The only test mode effect we found was a difference in completion time between PPTs and CBTs, and between CBTs with different levels of flexibility. The PPTs and the CBT modes with flexibility took longer to complete than the CBT modes without flexibility. It is interesting to note that the difference in completion times between the two conditions in Experiment 1 (3.8 minutes) is very close to the difference in completion times between the all-at-once, scroll condition and the one-at-a-time, no revisit condition of Experiment 2 (4.3 minutes).

Bocij and Greasley (2003) compared PPT and CBT formats and found that students felt CBTs were faster because they did not have to spend time writing down their responses. In the present study, however, this explanation does not hold up because the differences in flexibility among CBT formats produced similar completion time differences in Experiment 2. Instead, reviewing and changing answers probably explains most of the time difference between these test interfaces. Three to five minutes seems excessive in terms of accounting for the difference between writing letters vs. moving and clicking a computer mouse on a 30-item test. Instead, that time was more likely used to check over the answers. Increased flexibility in test format did not lead to increased performance, despite previous experimental evidence that changing answers improves scores (Kruger, Wirtz, & Miller, in review; Mueller & Wasser, 1977; Schwarz, McMorris, & DeMers, 1991; Vispoel, 1998).

It should be noted that although the tests used in this study were timed, they were not speed tests. Students were expected to be able to complete all the items on the test within the allotted time. We observed that students did so easily and did not feel rushed. Few studies were found that examined time differences for this type of test, despite the fact that timed, non-speed tests are commonly administered in the classroom. Our results provide some evidence that computerizing this format of test will not affect students' scores, even if the only available computer interface does not provide the same amount of flexibility as pencil and paper.

Though the observed time differences were statistically significant, it is difficult to assess whether they are practically significant. In both cases, the observed difference was approximately four minutes, or just over 10% of the total testing time. In most cases, four minutes is probably not a practical difference. Ten percent of total testing time might be practically significant for longer tests, but these results cannot be generalized to such a case.

The generalizability of the present study is limited by two factors. First, we used a sample of undergraduate students who we assumed to be computer-literate. However, the present study is consistent with other studies that found no test mode effect (Mason, Patry, & Bernstein, 2001; Mead & Drasgow, 1993; Neuman, & Baydoun, 1998; Ward, Hooper, & Hannafin, 1989; in one case for populations other than undergraduates Olsen, Maynes, Slawson, & Ho, 1989). Second, our test construction presented the exact same items in exactly the same order for all interfaces, so we did not exercise one of the main advantages of computerizing tests. It is unknown whether varying the order in which questions are presented to students would create a test mode effect.

In summary, the present study found a time advantage for test modes that do not provide flexibility. We found no difference in scores, regardless of the level of flexibility. These results provide evidence that teachers can attain some of the benefits of computerizing their tests, using whatever software is available and easy to use, without adversely affecting student performance. This is especially relevant for the public school environment, where 67% of teachers reported lack of training as a barrier to implementing computer use in the classroom (U.S. Department of Education, 2000). In the best case, test times might even be shortened.

REFERENCES

- Bocij, C., & Greasley, A. (1999). Can computer-based testing achieve quality and efficiency in assessment? *International Journal of Educational Technology*, 1, n1. Available: <http://www.ao.uiuc.edu/ijet/v1n1/bocij/index.html>.
- Bugbee, Jr., A. C., & Bernt, F. M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education*, 23, 87-101.
- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, 33, 593-602.
- DeAngelis, S. (2000). Equivalency of computer-based and paper-and-pencil testing. *Journal of Allied Health*, 29, 161-164.
- Dimock, P. H. (1991). The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement and Evaluation in Counseling and Development*, 24(3), 119-127.
- Frederico, P. A. (1989). *Computer-based and paper-based measurement of recognition performance*. (Navy Personnel Research and Development Center Report No. NPRDC-TR-89-7). San Diego, CA. (ERIC Document Reproduction Service No. ED306308.)
- Graud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement*, 10, 23-34.
- Inouye, D. K., & Bunderson, C. V. (1986). Four generations of computerized test administration. *Machine-Mediated Learning*, 1, 355-371.

Kruger, J., Wirtz, D., & Miller, D. T. (in review). Counterfactual thinking and the first instinct fallacy. *Journal of Personality and Social Psychology*.

Lee, J., Moreno, K. E., & Sympson, J. B. (1986). The effects of mode of test administration on test performance. *Educational and Psychological Measurement*, 46, 467-473.

Mason, B. J., Patry, M., & Bernstein, D. J. (2001). An examination of the equivalence between non-adaptive computer-based and traditional testing. *Journal of Educational Computing Research*, 24, 29-39.

Mazzeo, J., Druesne, B., Raffield, P. C., Checketts, K. T., & Muelstein, A. (1991). Comparability of computer and paper-and-pencil scores for two CLEP general examinations. *College Board Report No. 91-5*. New York. (ERIC Document Reproduction Service No. ED344902).

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin*, 114, 449-458.

Mills, C. N., & Stocking, M. L. (1996). Practical issues in large-scale computerized adaptive testing. *Applied Measurement in Education*, 9, 287-304.

Mueller, D. J., & Wasser, V. (1977). Implications of changing answers on objective test items. *Journal of Educational Measurement*, 4, 9-13.

Neuman, G., & Baydoun, R. (1998). Computerization of paper-and-pencil tests: When are they equivalent? *Applied Psychological Measurement* 22, 71-83.

Olsen, J. B., Maynes, D. D., Slawson, D., & Ho, K. (1989). Comparison of paper-administered, computer-administered and computerized adaptive achievement tests. *Journal of Educational Computing Research*, 5, 311-326.

Pomplun, M., Frey, S., & Becker, D. F. (2002). The score equivalence of paper-and-pencil and computerized versions of a speeded test of reading comprehension. *Educational and Psychological Measurement*, 62, 337-354.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*, 7, 20.

Schwarz, S. P., McMorris, R. F., & DeMers, L. P. (1991). Reasons for changing answers: An evaluation using personal interviews. *Journal of Educational Measurement*, 28, 163-171.

Spray, J. A., Ackerman, T. A., Reckase, M. D., & Carlson, J. E. (1989). Effect of the medium of item presentation on examinee performance and item characteristics. *Journal of Educational Measurement*, 26, 261-271.

Trotter, A. (2001). Testing firms see future market in online assessment. *Education Week on the Web*, 20(4), 6.

U.S. Department of Education National Center for Education Statistics. (2000). *Teachers' tools for the 21st century: A report on teachers' use of technology*. (NCES 2000-102).

Vispoel, W. P. (1998). Reviewing and changing answers on computer-adaptive and self-adaptive vocabulary tests. *Journal of Educational Measurement*, 35, 328-345.

Ward, T. J., Jr., Hooper, S. R., & Hannafin, K. M. (1989). The effect of computerized tests on the performance and attitudes of college students. *Journal of Educational Computing Research*, 5, 327-333.

Weinberg, A. (2001). Comparaison de deux versions d'une test de classement: Version papier-crayon et version informatisee [Comparison of two versions of a placement

test: Paper-pencil version and computer-based version]. *Canadian Modern Language Review*, 57, 607-627.

Zandvliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computers in Education*, 29, 423-439.

Direct reprint requests to:

Dr. Shawn M. Bodmann
Department of Psychology
University of Wisconsin-Madison
W. J. Brogden Hall
1202 West Johnson Street
Madison, WI 53706-1696
e-mail: smbodmann@wisc.edu