УДК 37.015.3:004

COMPUTER-BASED TESTING: ADVANTAGES AND DISADVANTAGES

I. P.Kuzmina
teacher of English
National Technical University of Ukraine "Kiev Polytechnic Institute"

У статті розкриваються переваги та недоліки комп'ютерного тестування.

Ключові слова: комп'ютерне тестування, інформаційні технології.

A Computer-Based Assessment, also known as Computer-Based Testing, e-exam, computerized testing and computer-administered testing, is a method of administering tests in which the responses are electronically recorded, assessed, or both. As the name implies, Computer-Based Assessment makes use of a computer or an equivalent electronic device (i.e. handheld computer). Computer-Based Assessment enables educators and trainers to author, schedule, deliver, and report on surveys, quizzes, tests and exams [*2*].

Computer-Based Testing may be a standalone system or a part of a virtual learning environment, possibly accessed via the World Wide Web.

The use of computers for testing purposes has a history spanning more than 20 years. In early studies, the main research focus was on whether computer-based tests were equivalent to paper-and-pencil tests when computers gave exactly the same tests as those given in paper-and-pencil formats. In order to define score equivalence, the American Psychological Association (APA) in 1986 published the Guidelines for Computer-Based Tests and Interpretations. The guidelines define the score equivalence of computerized tests and conventional paper-and-pencil tests as 1) the rank orders of scores of individuals tested in alternative modes closely approximating each other and 2) the means, dispersions, and shapes of the score distributions being approximately the same, or capable of being made approximately the same by rescaling the scores from the computer tests versions [*7*]. The guidelines also require that any effects due to computer administration be either eliminated or accounted for in interpreting scores. In their empirical study, Olsen *et al*. [*13*] compared paper-administered, computer-administered, and computer-adaptive tests by giving third- and sixth-grade students mathematics applications achievement tests. This study found no significant difference between paper-administered and computer-administered tests, and equivalences among the three test administrations in terms of score rank order, means, dispersions, and distribution shapes.

Mazzeo and Harvey [*12*] pointed out that computer-based test graphics may effect test scores and consequently their equivalence with paper-and-pencil versions, and that tests with reading passages may be more difficult when given on computers. Bunderson *et al*. [*6*] suggested performance on some item types such as paragraph comprehension are likely to be slower if presented by computer, while some types such as coding speed items are likely to be faster.

In reviewing all above-mentioned studies, Bugbee [*5*] concluded that the use of computers indeed affects testing; however, computer-based and paper-and-pencil tests can be equivalent provided the test developers take responsibility for showing that they are. Bugbee stated that the barriers to the use of computer-based testing are inadequate test preparation and failure to grasp the unique requirements for implementing and maintaining computer tests. In other words, Bugbee reminded us that some factors such as the design, development, administration and user characteristics must be taken into consideration when computers are used for testing.

As computer-assisted instruction (CAI) has grown in popularity, computer-based testing has become more and more appropriate for assessing students' CAI learning achievement. As Bugbee states [*5*], if what is being tested is done on or learned from a computer, then it is more appropriate to assess it by computer. Thus, computers are used as the sole vehicles for distributing tests, not only as alternatives to paper-and-pencil testing. Alessi and Trollip [*1*], in their classic book on computer-based instruction, devoted a chapter to the design, development, and use of computer-based testing. They pointed out that the two main ways of incorporating computers into the testing process are for constructing or administering tests. When constructing tests, test developers use computers' word processing abilities to write test items and use their storage capacities to bank and later retrieve test items. Jacob and Chase [*8*] pointed out that computers can present test materials paper-and-pencil test cannot, for example, 3-D diagrams in computer graphics, motion effects, rotating geometric forms, animated trajectories of rapidly-moving

objects, and plants seen from different angles. Shavelson *et al*. [*16*] further suggest using computer simulations for hands-on performance assessment. In their project “Electric Mysteries”, students were required to replicate circuits by manipulating icons of batteries, bulbs, and wires presented on a Macintosh computer.

When administering tests, computers can be used to provide individualized testing environments, that is, allowing students to take tests when they are ready. Moreover, test contents can be customized for students by providing different difficulty levels and emphases [*1*]. Computer-based testing can also be designed to provide test-takers with immediate feedback and scoring. However, Wise and Plake [*17*] found that immediate feedback may contribute to students’ test anxiety. Bernt *et al*. [*3*] also pointed out that general computer-test anxiety may influence test-takers. They considered that, although anxiety tends to be a random variable among people, it must be identified and dealt with. Jacob and Chase [*8*] also suggested discontinuing item-by-item feedback until further research has been done on the computer-test-anxiety issue.

Advancements in computer networking technology have allowed stand-alone computers to be equipped with powerful communication abilities, thus providing an alternative for assessing students’ learning achievements and attitudes. Students dispersed at distant sites may have options to take the test at different test locations and times. In addition to the traditional multiple-choice, fill in the blank, and short essay type questions, Rasmussen *et al*. [*14*] suggested Web-based instruction include participation in group discussions and portfolio development to evaluate students’ progress. Khan [*9*] also suggested Web-based instruction designers have facilities that allow students to submit comments about courseware design and delivery.

Although many researchers, e.g., Rasmussen *et al*. [*14*], J. Ravitz [*15*], considered testing and evaluation to be of utmost importance in Web-based instruction and suggested some design strategies and techniques, few usable systems have been developed and no empirical data collected to explore the feasibility of computer-assisted testing and evaluation on the Web. The search for creative and effective tools and methods for conducting testing and evaluation in such a complicated technology-dependent learning environment represents a challenge for system designers and instructional designers.

The advantages of administering tests by computer are well-known and documented, and include: 1) reduced testing time; 2) increased test security; 3) provision of instant scoring (the test can be discussed while the whole thing is fresh in the subject’s mind; in selection where the number of candidates again immediate results are valuable; where a huge number of subjects is tested this facility is not so important); 4) better use of professional time; 5) reduced time lag; 6) greater availability: individuals can be tested in a computer setting individually or in groups, usually in more user-friendly environments than the large classroom auditoriums where p-p tests have been administered traditionally. The computer format is also much more flexible than the printed page: for example, split screens could show stimuli such as a picture, as well as the possible responses. In addition, the computer format allows each examinee to work at his or her own pace, much more so than the p-p version; 7) greater accuracy: computers can combine a variety of data according to specific rules; human are less accurate and less consistent when they attempt to do this. Computers can handle extensive amounts of normative data, but humans are limited. Computers can use very complex ways of combining and scoring data, whereas most humans are quite limited in these capabilities. Computers can be programmed so that they continuously update the norms, predictive regression equation, etc., as each new case is entered; 8) greater standardization: the computer demands a high degree of standardization both test procedures and test interpretations, and, ordinarily, does not tolerate deviance from such standardization; 9) greater control: this relates to the previous point, but the issue here is that the error variance attributable to the examiner is greatly reduced if not totally eliminated; 10) greater utility with special students and groups: there are obvious benefits with computerized testing of special groups, such as the severely disabled, for whom p-p tests may be quite limited or inappropriate; 11) long-term cost savings: although the initial costs of purchasing computer equipment, of developing program software, etc., can be quite high, once a test is automated it can be administered repeatedly at little extra cost; 12) easier adaptive testing: this approach requires a computer and can result in a test that is substantially shorter and, therefore, more economical of time. The test can also be individualized for the specific examinee.

In reading the above list, you may conclude that some advantages may not really be advantages. There may be some empirical support for this, but relatively little work has been done on these issues. For example, immediate feedback would typically be seen as desirable. S. L. Wise and L. A. Wise (1987) compared a p-p version and two of versions of a classroom achievement test with third and fourth graders. One of versions provided immediate item feedback and one did not. All three versions were equivalent to each other in mean scores. However, high math-

achievers who were administered the cf version with immediate feedback showed significantly higher state anxiety; the authors recommended that such feedback not be used until its effects are better understood [*11*].

In addition to the disadvantages just discussed, we might consider that computerized testing reduces the potential for observing the subject's behavior. As we have seen, one of the major advantages of tests is that subjects are presented with a set of standardized stimuli, and the examiner can observe directly or indirectly the rich individual differences in human behavior that enhance the more objective interpretation of test results. With computerized testing, such behavior observation is severely limited [*11*].

One more disadvantage: the need for individual computer terminals for each person limits the number of subject who can be tested at any one time.

Let's point out the important aspects of computerized testing:

1. *Items*. There is no magic about computer testing. A computerized test is no more or no less than the sum of its items, as is the case with traditional tests. However, it is possible, in principle, to use items that could not be presented other than by computer. An obvious example arises in the sphere of tests of reaction time and tracking tasks, such as found in arcade computer games. However, a computer test, even if it consists of what might be called computer bound items, must still be judged against the standard psychometric criteria of reliability, discriminatory power, validity and quality of normative data, where these are applicable.

2. *Comparability between a paper and pencil test and a computer-administered test.* It is possible to computerize virtually any traditional test. It is far easier to present on the computer screen verbal and numerical items than visual items where there is always the possibility that the screen image will be different from the printed test, even with modern graphics and light-sensitive pens. Nevertheless, no matter how identical the two tests appear to be it is essential that the reliability, validity and standardization of the computer version be checked. Furthermore, it is essential to show that the correlation between the two versions is high. Indeed, if the computer version is to be regarded as identical with the traditional test this correlation should be at least .9. Thus the computer test should be considered to be a parallel form. Generally, it must be said, as Bartram and Bayliss (1984) point out, computer-administered tests and their traditional counterparts have turned out to be highly equivalent.

There is obviously a severe problem here. If, due perhaps to the low reliability of the original test, this correlation is only around .5, then it is impossible to regard the two tests as measuring the same variable. Clearly only tests with high reliability should be transmuted for computer. In any case, as has been said, new reliability, validity and standardization data should be collected. Comparability has to be demonstrated rather than assumed.

3. *Computer test instructions.* In a traditional test it is essential that the instructions are comprehensible to all subjects. In a computer-presented test it is similarly absolutely essential that the procedures for answering the questions, for obtaining the next question, for altering responses and for looking back (if it is allowed) are clear and easily worked by subjects. Computerized tests must be user friendly in the simplest sense. If subjects are anxious about working the machine or are making errors as they proceed, or are unable to operate the computer, the test will fail.

4. *Indices of item difficulty.* These, or other similar indices, can be stored in the computer. This allows the tester to present a sample of the items in the test and yet arrive at an accurate score. This is known as tailored testing.

5. *Analysis of data.* An enormous advantage of computerized tests is that data analysis, both for individuals and for groups, is made absurdly easy.

a. *Individual data*. The computer can automatically store the results of the test item by item, as well as any other relevant information. Before starting the test all subjects should be required to insert the following information, as a minimum: age, sex and level of education (in numerical form: for example, 1 for no qualifications up to 5 for a higher degree). This means that the computer can immediately produce the subject's score and its standard error, and the most appropriate standard score, if norms are established for the computer test. In addition it can show items which are wrong, or, in the case of personality and attitude tests, items not endorsed in the keyed direction, all of which may be useful information for the tester, in various applied settings. In vocational guidance, for example, it is often valuable to discuss actual responses to individual items with the subject. Thus the computer can immediately, on completion of the test, provide the raw and standardized test score and any appropriate standard errors.

b. *Analysis of data*. The computer stores the results of each subject's data. Thus, after a substantial number of subjects has been tested, it is simple to analyse the data. Item analysis, factor analysis, group norms, comparisons across categories of subjects by analysis of variance, are all possible with commercially available programmes.

6. *Presentation of results to subjects.* Immediately the test is finished the computer can present the results to the subject, either on screen or as a printed document [*10*].

Incidentally it should be pointed out that some of these facilities are possible with paper-and-pencil tests which are computer scored. Here the test is administered to subjects in the usual way, but the responses are punched into the computer. This allows the printed report for the subjects and comparisons with norm groups to be produced. It also allows a database to be built up for the development of special norms. What of course is not possible is the presentation of items appropriate to the subject, as determined by the subject's responses.

For now, much effort has been devoted to "translating" p-p versions to cf versions, and relatively little effort has been devoted to creating new computer-administered tests. A number of tests have however been developed specifically for computerized use, and some of these take advantage of the graphic possibilities of the computer.

A number of studies have investigated the use of light pens, joysticks, and other mechanical-electronic means of responding to test items for disabled individuals who are not able to respond to tests in traditional ways. The results suggest substantial equivalence across response modes (e.g., Carr, Wilson, Ghosh, Ancill, Woods, Ridgway, MacCulloch).

Computers can easily assess response time, that is, how fast a subject responds. Response time (or reaction time, response latency) to questionnaire items could be a useful additional measure in a number of research areas (Ryman, Naitoh, Englund, et al.). Some authors, for example, have argued that such latencies can potentially be indicative of meaningful variables. On a personality test, longer latencies may reflect more "emotional" items (Space).

As we can see there certain advantages to computer-administered and computer-scored tests – especially the rapid calculation of a subject's results and the immediate presentation of her or his scores in terms of normative groups or other criteria. In addition there are advantages in the ability to present subsets of items. There are further advantages including the ability to store all results and develop new or local norms, and the opportunity they allow the tester to examine the statistical quantities of the test, right down to the item level. Finally, types of item can be used which are impossible in the traditional test.

All this is good and provided that the ethical problems (of presenting results to subjects without their being able to discuss their implications and their own reactions to them) are dealt with, computer-administered tests can be useful. Much, in the end, depends upon the humanity and awareness of the particular psychometrist. There is little doubt, however, that computer-administered testing can lead to ethical abuse.

As with any technological revolution, there are problems and challenges to be met, some of which have been discussed above, some of which have been ignored because they would take us too far afield, and some of which we are not even as yet aware.

REFERENCES

1. *Alessi S. M., Trollip S. R.* Computer-Based Instruction: Methods and Development. – Englewood Cliffs, NJ: Prentice-Hall, 1991. – P. 205-243.
2. *Asuni N.* TCExam: Computer-Based Assessment. – http://www.tcexam.org
3. *Bernt F.M., Bugbee A.C., Arceo R.D.* Factors influencing student resistance to computer administered testing // J. Res. Comput. Educ*.* – 1990. – Vol. 22. – № 3. – P. 265-275.
4. *Bubgee A.C.* Examination on Demand: Findings in Ten Years of Testing by Computer 1982-1991. – Edina, MN: TRO Learning, 1992.
5. *Bubgee A.C.* The equivalence of paper-and-pencil and computer-based testing // *J. Res. Comput. Educ*. – 1996. – Vol. 28. – № 3. – P. 282-299.
6. *Bunderson C. V., Inouye D. K., Olsen J. B.* The four generations of computerized educational measurement // Educational Measurement. – New York, NY: Amer. Council Educ. – Macmillan, 1989. – P. 367-407.
7. *Guidelines for Computer-Based Tests and Interpretations*. – Washington, DC: Amer. Psychol. Assoc., 1986.
8. *Jacobs L. C., Chase C. I.* Developing and Using Tests Effectively: A Guide for Faculty. – San Francisco, CA: Jossey-Bass, 1992. – P. 168-177.
9. *Kline P.* The handbook of psychological testing. – 2000. – P. 97-106.
10. *Khan B.H.* Web-Based Instruction (WBI): What is it and why is it? / B. H. Khan // Web-Based Instruction. – Englewood Cliffs, NJ: Educational Technology, 1997. – P. 5-18.
11. *Marla L. Domino*. Psychological testing: an introduction. – 2006. – P. 475-485.

12. *Mazzeo J., Harvey A. L.* The Equivalence of Scores from Automated and Conventional Educational and Psychological Tests. – College Entrance Examination Board, New York, College Board Rep. 88-8, 1988.
13. *Olsen J. B., Maynes D. D., Slawson D., Ho K.* Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. – San Francisco, CA, 1986.
14. *Rasmussen K., Northrup P., Lee R.* Implementing Web-based instruction // Web-Based Instruction. – NJ: Educational Technology, 1997. – P. 341-346.
15. *Ravitz J.* Evaluating learning networks: A special challenge for Web-based instruction. – NJ: Educational Technology, 1997. – P. 361-368.
16. *Shavelson R. J., Baxter G. P., Pine J.* Performance assessments: Political rhetoric and measurement reality // Educ. Res. – 1992. – Vol. 21. – №. 4. – P. 22-27.
17. *Wise S. L., Plake B. S.* Research on the effects of administering tests via computers // Educ. Meas.: Issues Practice. – 1989. – Vol. 8. – № 3. – P. 5-10.

Стаття надійшла до редакції 31.03.2010 р.