ED 282 897

AUTHOR
TITLE

PUB DATE
NOTE

Testing and Computer-Based Instruction: Psychometric Considerations

Paul D. Sarvela, Ph.D. & John V. Noonan, Ph.D.

Department of Health Education
College of Education
Southern Illinois University
Carbondale, IL 62901

Ford Aerospace & Communications Corporation
Western Development Laboratories Division
7100 Standard Drive
Hanover, MD 21076

Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., April, 1987.

Testing and Computer-Based Instruction: Psychometric Considerations

Abstract

Computers are used in a number of ways to aid in the design, development, and delivery of tests in computer-based instruction (CBI) settings. Although there are many advantages to using computer-based tests (CBTs) linked to CBI, there are also several difficulties associated with their use. One major problem related to the use of CBTs is that in certain instructional settings, it is difficult to conduct psychometric analyses of the test results. This paper examines several measurement issues which surface when CBT programs are linked to CBI, including CBT standards, decisions on item types, the contamination of items, and non-equivalence of groups.

Introduction

Computers are used in a number of ways to aid in the design, development, and delivery of tests in computer-based instruction (CBI) settings. In addition to the use of computers in the delivery of "traditional" forms of tests (such as a test composed of multiple-choice test items) computers can also be used to simulate scenarios which require student demonstration of complex cognitive and psychomotor problem-solving skills, such as the evaluation of pilots in a flight simulation (e.g., Breidenbach & Frank, 1984; Conkright, 1982; Williams, 1984) and the simulation of complex medical-related problems when examining medical student competencies (i.e., Norman, Muzzin, Williams, & Swanson, 1985).

Although there are many advantages to using computer-based tests (CBTs) linked to CBI, there are also several difficulties associated with their use. One major problem, in certain CBI settings, is the difficulty in conducting psychometric analyses of the test results. Discontinue criteria, random selection of items, individualized instruction (which impacts treatment effects and the calculation of gain scores), and embedded test strategies are some of the many "advantages" of CBI which create a number of problems for the CBT measurement specialist.

The purpose of this paper is to describe several measurement problems associated with the use of CBT programs when they are a part of a larger CBI curriculum. Specifically, this paper examines CBT standards, decisions of item types, the contamination of items that arise from certain test design strategies, and the non-equivalence of comparison groups in item analyses.

Computer-Based Testing Standards

The Committee on Professional Standards and the Committee on Psychological Tests and Assessment of the American Psychological Association have developed a set of preliminary guidelines on the use of computer-based tests and their resulting interpretations (APA, 1986). Specific recommendations are outlined for those individuals who build CBT software. One guideline, which has important implications for CBT design, refers to the human factors component of CBT development. The guidelines state that:

> computerized administration normally should provide test takers with at least the same degree of feedback and editorial control regarding their responses that they would experience in traditional testing formats (APA, 1986, p. 12).

These testing recommendations have interesting implications for the CBT developer. For example, if an examinee can change answers (as in paper-pencil test), when is the answer to be logged onto the record file or data tape? If answer changing is allowed, it is difficult to use adaptive testing because item presentation is dependent upon previous responses. Conversely, an inability to change a response to an item can create other problems. If an examinee needs to change an answer, either because he feels another selection is more appropriate, or because he made a keyboard error (accidentally pressed down the wrong key), he should be allowed to change the item. An inability to change items can be unfair to examinees and could affect the reliability and validity of the test.

In addition to the computer-specific human-factors issues which must be considered when designing CBTs, the psychometric analysis must be considered. Measurement standards which apply to traditional forms of tests apply to CBTs as well. Therefore, information concerning reliability, validity, item analysis, and norms should be gathered as part of the CBT development process.

Item Type

Different item types are normally used to test different types of learning. For example, if a tested objective is classified as a "recall-fact" learning objective, it should only be tested by a constructed-response item type (e.g., fill-ins, short answer, essay). This is because "recall-fact" information must be memorized, and a constructed-response item is the only item that will theoretically measure "recall-fact" learning (Wulfeck, Ellis, Richards, Wood, & Merrill, 1978). Selected-response items (e.g., multiple-choice, matching, true/false) require only recognition of the answer, not total recall. Therefore, if rigorous standards are emphasized in the test specifications, only constructed-response items would be acceptable methods of testing "recall-fact" objectives. Problems arise, however, when constructed-response items are designed and developed for the computer. Constructed responses require considerably more complex analyses. Unless the computer has a natural language processing capability, it becomes nearly impossible to program all the possible correct answers for a short-answer item, when considering alternate wording, spacing, spelling errors, and alternate correct answers. The first difficulty arises in trying to detail all possible correct answers. As an example, consider the following constructed-response item: "What are the two steps in preparing the XYZ radio tuner?" Suppose that the two steps are: (1) turning the power on and (2) turning the mode selector dial to "tune." Further suppose that the order of these steps is not important. Following are some correct answers:

a. Turn it on and turn the mode dial to tune
b. Set mode switch to Tune and then turn the power on
c. First you press the power switch, then you rotate the other dial to "tune."
d. I think you flip the power switch and turn the dial selector to tune.

The list could obvously go on ad infinitum.

The second problem is in programming time. Without some kind of artificial intelligence capability, a tremendous about of programming is involved for even a partial subset of all possible correct answers.

There are psychometric implications as well. Students could supply correct answers that simply are not recognized by the computer; the result could be lower reliability and poorer discrimination indices.

Because of the above-mentioned problems, a practical compromise would be to use only selected-response items on CBTs. The design and development process for selected-response items is much quicker, and the response analysis is more accurate. CBT technology is simply not well-prepared to handle constructed-response items at this time.

Contamination

Another issue facing psychometricians attempting to validate tests in CBT environments is the problem of item contamination. By using the instructional design capabilities of CBI systems, it is possible to allow students to preview test items, receive feedback on the correctness of their answers while items are still being presented, or, retake items which were drawn randomly from an item pool.

When students are allowed to preview items on a test (since some CBI programs allow students to freely move in and out of tests) a major contamination problem occurs. In this situation, one risks having students memorize test items, and not learn the total domain of knowledge to be taught. One must decide if testing the total domain of knowledge is important, or, if an understanding of specific test items is important. If understanding the domain of learning is critical, and the test items are drawn from a pool of possible items, then a preview of test items is not recommended. Conversely, if an understanding of discrete facts is crucial, and no sampling is done, then a preview of the test items should present no problems in the interpretation of the test.

Another problem related to contamination is test-item feedback while the test is being taken. A major advantage of CBI is the capability of immediate scoring and feedback. However, this capability is not always recommended for testing. If a student receives feedback after each item, items which are dependent upon each other (i.e., an item which requires the student to use the result from item 3 to compute item 4) would be contaminated. Or, the correct answer for one item could provide subtle clues to the correct answer on another item. There are motivational concerns as well. If a student is consistently answering items incorrectly, the negative feedback might be detrimental to motivation on future items. Likewise, a series of correct-answer feedbacks can promote greater motivation in future items. The danger is in the differential effects of item feedback across high and low achieving students. Test administrators are usually cautioned about giving item feedback during the test's administration. In addition, test directions often caution about the dangers of giving subtle cues about the correctness of the student's response (i.e., Wechsler, 1974).

One final contamination problem results from the practice of selecting items randomly from an item bank for a particular test. Computers allow us to develop large item pools and then apply various sampling strategies for arriving at the particular subset of items that a given student will see (e.g., random without replacement, same items in shuffled order.). If a student fails a test, and is then rerouted through the lesson, he is usually retested on the same material. When items are selected randomly from a pool, there is a possibility that the student might see the same items on a second or third try (the probability of seeing an item on a retry is related to the size of the item pool - the larger the pool, the lower the probability of seeing an item). Because of this problem, it may be a better practice to use a sequential method of presenting the items than a random presentation of items. This would eliminate the risk of the student seeing the same item twice. It should also be noted that this problem is exacerbated when item feedback is given. If item feedback is provided, second attempts at tests should logically contain new items.

Non-Equivalence of Groups

A final area of CBT psychometric difficulty relates to the non-equivalence of groups. Because of the unique nature of CBT (Noonan & Sarvela, 1987), it often occurs that for a given test, different students see:

1) a different number of items
2) different items
3) a different item order

4) items at different times in the course

These problems occur when test items are drawn randomly from an item pool, when the item order is mixed by the test designer, when discontinue rules are used, and when the test designer allows free access to pretests and posttests.

The net result is that evaluation of tests is thwarted by the non-equivalence of any comparison groups. The central problem is that when tests have the above-mentioned characteristics, there is no sensible total test score upon which to base frequency distributions or item analyses. Consider a situation where a given posttest has 20 items in an item pool. Ten items will be selected at random for presentation, and the test will be discontinued once the student has passed 7 items or failed 4 items (cut score of 7). The argument could be made that there is no reason to have students continue to take items when the mastery-nonmastery decision has already been made. In this situation, it would be difficult to compute a total test score, since the maximum correct is 7, and students can achieve the score of 7 in 7, 8, 9 or 10 items.

Moreover, the items themselves differ, due to their random selection from a 20-item pool. There are at least two serious problems with random item selection. The first problem is that there is an implicit assumption that the items administered to one student will be equal in difficulty to items that are presented to another student. For example, imagine that a pool of items has an average p-value (difficulty index) of .80 and a standard deviation of p-values of .12. If the test is going to be fair to students, the items that one student sees should be comparable in difficulty to the items which another student sees. In the long term, random selection will produce comparable tests, but one certainly would expect that at times one student would receive all of the easier items and another would receive the harder items. The frequency with which this occurs would depend upon the degree of variance in item difficulty. One possible control for this undesirable effect would be randomly select items within strata of difficulty. For example, one item could be randomly selected from the p-value range of .90-1.00, three items from the range of .80-.89, and one item from the range of .00-.79.

The second conceptual difficulty with random item selection relates to compromises on program and test evaluation. If students see different items, it becomes extremely difficult to compute item and test statistics (e.g., total score, point biserial, KR-20). The problem is that there is no sensible total score. With random item selection, a total test score only becomes defensible for item analysis if every item is of equal difficulty and equal discrimination (otherwise, the students have not seen the "same test"). Further, pretest and posttest comparisons presume parallel forms of a test (equal means, standard deviations, item inter-correlations, reliabilities, and validity coefficients). As with the problem of total test score statistic computation, with random item selection, parallel test criteria can only be met if each item in the test domain pool is of equal difficulty and discrimination, a highly improbable condition. (It is important to note that item-response theory provides a way of handling these difficulties, but the solutions require estimates of item parameters that can only be obtained with large samples. Courseware development efforts often do not have access to large samples for pilot testing.)

Many of the above-mentioned problems disappear if items are presented in sequence. Usually, in a sequential item delivery CBT strategy, a set number of items are presented in a particular order. (This format is most closely analogous to a paper-pencil test.) Total test scores fit well into the logic of test theory and less concern can be given to establishing equal item difficulty and discrimination.

Summary

To conclude, computers are currently used in a number of ways to aid in the design, development, and delivery of tests in CBI settings. Although CBTs can be used effectively when linked to CBI, there are several difficulties associated with their use. This paper has described several problems concerning the use of CBT programs. The discussion has focused on problems related to CBT standards, item type, item contamination, and non-equivalance of groups. Some possible solutions to these problems have been offered.

Acknowledgments

The authors would like to thank Dr. William Sweeters, of Ford Aerospace and Communications Corporation, and Ms. Joyce V. Fetro, of Southern Illinois University, for their helpful comments on an earlier draft of this manuscript.

References

American Psychological Association (APA). (1986). Guidelines for Computer-Based Tests and Interpretations. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Briedenbach, S.T., & Frank, L.H. (1984). The use of graphic performance feedback in pilot carrier landing training. Training Technology Journal, 1, 44-50.

Conkright, T.D. (1982). PLATO applications in the airline industry. Journal of Computer-Based Instruction, 8, 49-52.

Noonan, J.V., & Sarvela, P.D. (1987). Implementation Decisions in Computer-Based Testing Programs. Manuscript submitted for publication.

Norman, G.R., Muzzin, L.J., Williams, R.G., & Swanson, D.B. (1985). Simulation in health sciences education. Journal of Instructional Development, 8, 11-17.

Wechsler, D. (1974). Manual for the Wechsler Intelligence Scale for Children-Revised. San Antonio, Texas; The Psychological Corporation.

Williams, A. (1984). The application of computer-assisted instruction in F/A-18 pilot training. Training Technology Journal, 1, 44-50.

Wulfeck, W.H., Ellis, J.A., Richards, R.E., Wood, N.D., & Merrill, M.D. (1978). The Instructional Quality Inventory. (NPRDC SR-79-3). San Diego: Navy Personnel Research and Development Center.