11-1-1986

Do We Really Know What Makes Educational Software Effective? A Call for Empirical Research on Effectiveness

Karen Jolicoeur

Dale E. Berger
Claremont Graduate University

Do We Really Know What Makes Educational Software Effective? A Call for Empirical Research on Effectiveness

Karen Jolicoeur and Dale E. Berger

Educators who are responsible for selecting microcomputer software for classroom use are faced with a difficult task. The dramatic increase of microcomputers in schools in recent years has been mirrored by a tremendous proliferation of educational software. There are now more than 7,000 commercially produced educational software programs for microcomputers on the market (EPIE, 1986), and about 100 new programs are published each month ("New Releases," 1985-1986). This vast array of software leaves many educators in a quandary as how best to select the most appropriate programs for their classrooms. In this article we describe the current state of evaluation research with educational software, and discuss how popular software review methods fall short of meeting our need to know how well specific programs work.

Outcome Evaluations of Educational Software

Surprisingly, there have been very few research studies of the educational effectiveness of individual software programs for microcomputers. Research has shown that computer-assisted instruction (CAI) is an effective medium for improving academic skills in significantly less time than conventional classroom methods (Kulik, Bangert, and Williams, 1983; Kulik, Kulik, and Cohen, 1980; Thomas, 1979). However, these earlier studies were concerned with larger and older computer systems rather than with the microcomputer systems that are so widely available in classrooms today.

Karen Jolicoeur is a doctoral student and Dale E. Berger is Professor, Psychology Department, The Claremont Graduate School, Claremont, California. This research was supported in part by a grant from the Irvine Foundation to the Claremont Center for the Study of Pre-Collegiate Education.

Empirical information on specific factors that make educational software effective in reaching instructional objectives would be of considerable value. First, educators must know which factors to consider if they are to be successful in selecting appropriate and effective software for their classrooms. Second, software developers also need information on factors associated with effectiveness if they are to improve the quality of future educational software. Finally, those who review software need to know how specific program characteristics influence effectiveness in order to develop valid evaluation measures that accurately predict the educational value of software.

We discovered the dismal state of evaluation research on commercially available educational software for microcomputers when we attempted to conduct a meta-analysis on outcome studies. As a first step in the meta-analysis, we sought to obtain a comprehensive collection of studies that measured the effectiveness of commercially available CAI for microcomputers. A computerized literature search of ERIC, *Psychological Abstracts*, and *Dissertation Abstracts* through mid-1985 was conducted, as well as a search through a wide range of recent journals. In addition, about 200 letters were mailed to researchers and institutions known to be studying the process of educational computing, requesting information on published and unpublished outcome evaluation studies that used microcomputers and commercial CAI. We received 63 replies, including several warnings that we would not be able to find any such studies. From all sources we were able to compile a list of only 47 outcome studies. To be useful for our meta-analysis, a study had to meet three conditions: (1) the study must have measured the effects of an individual software program; (2) performance must have been measured by an objective test; and (3) there must have been a control group. Only two of the 47 studies (Davis, 1985; Watkins and Abram, 1985) met all requirements for the proposed meta-analysis. The most common reason a study had to be rejected was because several software programs were used concurrently. While a test of the simultaneous application of multiple CAI programs may provide general information regarding CAI effectiveness, it does not provide information that allows one to identify *specific characteristics* of software that are associated with effectiveness.

Obviously, the proposed meta-analysis is not possible on the current data base, since there have been very few studies of the actual effectiveness of educational software for microcomputers. Our next step was to examine the adequacy of published evaluations of software as measures of

effectiveness. These reviews are used by many educators to help with selection of software, with the expectation that the ratings are valid indicators of how well the software will work in the classroom (e.g., Schiffman, 1986).

Popular Methods for Reviewing Educational Software

Although there are many sources of educational software evaluations, the most extensive and widely used evaluations of these products are probably those provided by EPIE (Educational Products Information Exchange) and by Microsift (from the Northwest Regional Educational Laboratory). Since these two review services are among the best, they will serve well as examples. EPIE's software evaluation method is an adaptation of a successful textbook evaluation protocol that provides numeric ratings of overall instructional design features, technical software design features, and a summary recommendation (EPIE, 1985). In addition to these ratings, a "PRO/FILE" provides information pertaining to the reviewer's judgments of the program, recommendations to the software publisher, and content and management descriptions of the program. EPIE's PRO/FILE represents a composite of evaluations and judgments averaged across several certified reviewers, each of whom has received special training in completing EPIE's 16-page evaluation instrument. EPIE claims that their certification and evaluation process leads to high inter-rater reliability (EPIE, 1985).

In contrast to EPIE's 16-page evaluation instrument, Microsift uses a much shorter three-page software description and evaluation instrument. The Microsift evaluation process was tested over a 12-month period by educators from 26 educational institutions across the United States (Otte, 1984). The evaluation includes descriptive statements concerning the potential use of the program, as well as 21 specific reviewer agreement/disagreement ratings that are separated into three major areas of software components: content, instructional characteristics, and technical characteristics. The 21 factors are each rated on a four-point scale from "strongly disagree" to "strongly agree." The reviewers also provide numerical ratings of the software's content, instructional characteristics, technical characteristics, and an overall summary recommendation. Each software package reviewed by the Microsift method is evaluated by a minimum of three and a maximum of six reviewers (Otte, 1984). Like EPIE's final PRO/FILE report, the final Microsift Courseware Evaluation report is a composite of judgments and evaluations averaged across all reviewers.

On the surface, the evaluation procedures used by EPIE and Microsift appear intuitively sound. However, by psychometric standards both methods suffer badly. The summary evaluations of the programs are based primarily on the reviewers' subjective opinions rather than on operationally defined variables. For example, EPIE's evaluation form includes the following instructions:

> The following categories should be used in determining your overall ratings of the program. They are not exhaustive, but suggest the areas you should consider in making your judgments. All items may not be appropriate to use in rating all programs. Consider the scope and intended purpose of the program (e.g., it may be inappropriate to penalize a simulation that has no management system).
>
> | **Instructional Design** | **Software Design** |
> |---|---|
> | Goals and Objectives | Technical Quality |
> | Content | Graphics/Audio Quality |
> | Methods/Approach | User Control/Interactivity |
> | Documentation/Support Materials | Branching |
> | Evaluation/Tests | Management/Record keeping |
>
> Rate each of the following on a scale of 1 (lowest) to 10 (highest):
>
> INSTRUCTIONAL DESIGN ..... SOFTWARE DESIGN .....

Raters are likely to have a variety of opinions regarding what constitutes educational factors such as "goals and objectives." Thus, even if two evaluators rated the instructional design equal to 6, it is not clear exactly what qualities of the program elicited this rating.

Microsift's ratings of quality are subject to the same criticism. For example, raters are asked to complete the following quality judgments:

> Write a number from 1 (low) to 5 (high) which represents your judgment of the quality of the package in each division:
> ..... Content Characteristics
> ..... Instructional Characteristics
> ..... Technical Characteristics

Once again, it is not clear exactly how each rater defines content, instructional, and technical characteristics. Furthermore, since educators reading the reviews have their own subjective opinions regarding what constitutes high instructional and technical quality, their interpretations of the eval-

uations may be very different from what the raters intended. To assess the validity of these subjective judgments, we next compared reviews from EPIE and Microsift for the same set of software.

Validity of Current Software Reviews

In February, 1986 EPIE's on-line computer service, *The Educational Software Selector* (EPIE, 1986), listed 573 educational programs reviewed by EPIE and 238 programs reviewed by Microsift. The level of overall recommendation on a four-point scale was available from both EPIE and Microsift for 82 programs. A Pearson correlation (r) was calculated between the recommendations of EPIE and Microsift to determine how closely these evaluations agreed. The correlation was only .33, significantly different from zero ($p < .001$) but far below acceptable levels of reliability for alternate measures of the same concept (cf. Nunnally, 1978).

Although the overall recommendations were not reliable across the two evaluation services, one might expect higher reliability for ratings of more narrowly defined characteristics of the software. Ratings of the quality of instructional design characteristics and technical design characteristics were available from both evaluation services for 29 programs. Although each concept was not defined in exactly the same way by the two evaluation services, they overlap substantially. The observed correlation between EPIE and Microsift's overall recommendations for this set of programs was only .22, a value that statistically is not significantly different from zero! The two review services showed even weaker agreement on the other scales, with the ratings of instructional features correlating .13, and the ratings of technical features correlating .07 (neither significantly different from zero).

These low correlations suggest that the reviews may have very weak construct validity. Construct validity is the term used to describe how well an instrument measures what it was designed to measure. To demonstrate construct validity, two measures of the same construct must be highly correlated (showing convergent validity), and they should not correlate highly with measures of constructs which are intended to be different (showing discriminant validity). This premise was used by Campbell and Fiske (1959) in their multitrait-multimethod approach to assessing convergent and discriminant validity. The same logic can be applied to the current data, as shown in Table 1, by comparing the averages of two sets of correlations. The first set consists of correlations between measures of the same concept from the two evaluation services (e.g., the correlation between EPIE and Microsift's ratings of quality of instructional design characteristics for the same programs). These correlations are enclosed in circles in Table 1. The second set consists of correlations between different concepts using ratings from different services (e.g., the correlation between EPIE's rating of instructional characteristics and Microsift's rating of technical characteristics). These six correlations are enclosed in dashed triangles. The averages of the two sets of correlations were almost identical, .14 for the first set (same concepts) and .12 for the second set (different concepts).

Another criterion for discriminant validity proposed by Campbell and Fiske (1959) is that correlation between alternate measures of the same concept should be larger than correlations between measures designed to get at different concepts "which happen to employ the same method" (p. 83). However, in Table 1 we see that the correlation between ratings of instructional and technical characteristics was .54 for EPIE and .55 for Microsift. These correlations, which are enclosed in solid triangles, are both much larger than the average correlation of .14 between measures of the same concepts from the two review services. (We did not include correlations of the overall recommendation with the ratings of the instructional and technical characteristics, since the overall rating is based largely on these two characteristics.)

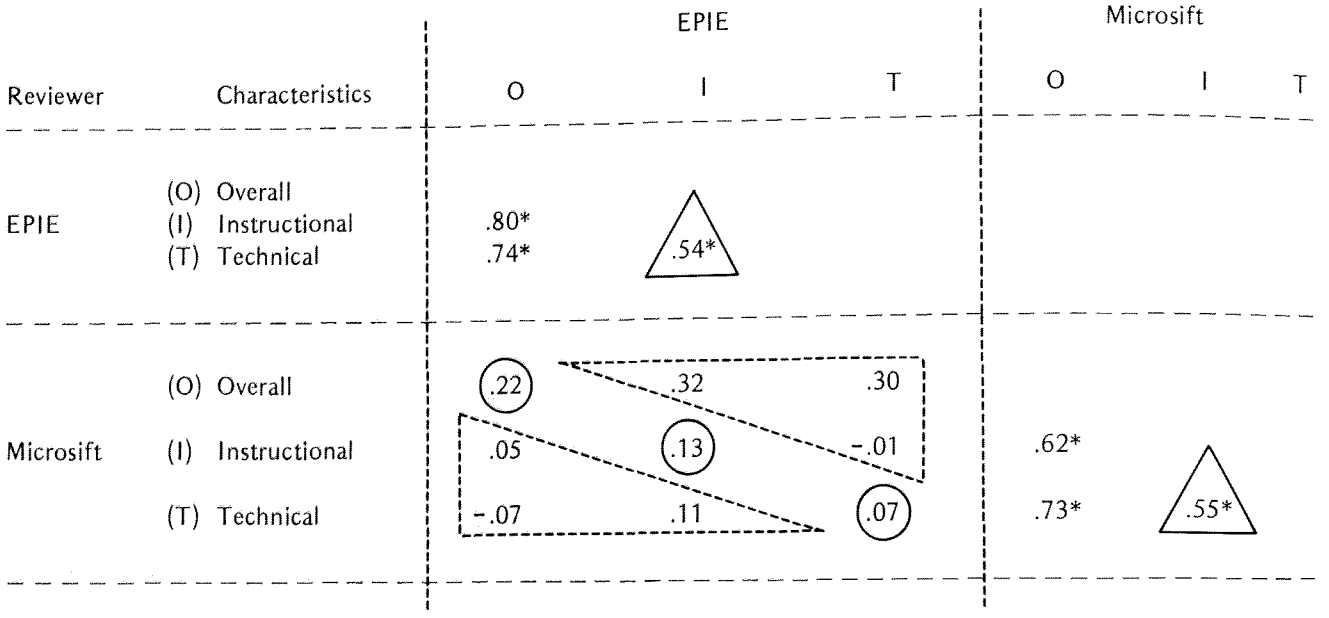
Implications for Educators

The unfortunate implication of these data is that there is little evidence for convergent validity and no evidence for discriminant validity of ratings in typical software reviews. Measures of closely related concepts from EPIE and Microsift have very weak correlations that are at the same level as correlations between measures of very different concepts from the two services. At the same time there is a strong "halo" effect whereby reviewers who rate a program high on instructional characteristics are likely also to rate the program high on technical characteristics.

The striking lack of agreement between EPIE and Microsift concerning the quality of specific educational software forces one to question the usefulness of such evaluations. While recommendations of experts would seem to be of some value, the low level of agreement between raters using similar systems reveals the danger of relying too heavily on subjective evaluations of effectiveness. Since the ratings do not agree with each other, they both clearly cannot be valid indicators of the actual effectiveness of the software. Of course, even if the evaluation services agreed, there would be no guarantee that the ratings would be valid

Table 1

Correlations of Software Ratings from EPIE and Microsift



Legend
○ = Same trait, different method
△ = Different trait, same method
⟍ = Different trait, different method

Note: Pearson correlations of ratings from 29 programs rated by both review services.
*$p < .01$

predictors of the effectiveness of the software. It should be noted that these criticisms are not specific to EPIE and Microsift, but they apply to any evaluations based on subjective judgments alone. The clear implication is that reviews currently are not able to provide educators with the basic information on program effectiveness that is so essential for appropriate software selection.

Conclusion

The only way to establish the validity of a system of evaluation for educational software is to demonstrate that highly rated programs do in fact teach academic objectives better and/or faster than lower rated programs. This means that controlled outcome studies are required, whereby gains in academic achievement attributed to the use of specific software can be measured with objective tests. However, as discussed earlier, there are very few empirical studies of instructional effectiveness. Consequently, we simply do not know which programs teach educational objectives better or faster than others.

Further, we do not know which specific factors contribute significantly toward making educational software more or less effective. While there is a rich literature in education and psychology to identify factors that have been shown to be important in other instructional contexts, the small amount of outcome evaluation research currently available on educational software is not adequate to validate these factors in the microcomputer context of instruction. Until a much larger number of outcome evaluations is available, it is not possible to develop a valid evaluation method that can predict the ability of a specific piece of computer software to attain educational objectives.

Finally, educators need to interpret published recommendations with caution. Ratings from different sources may not agree well with each other (they may have low reliability), and there is no evidence that subjective judgments of effectiveness are predictive of the actual effectiveness of programs (the validity of the ratings is unknown). Software reviews appearing in this magazine, unlike many other reviews, *are* based on actual tryouts with students in addition to subjective expert judgments.

There is a pressing need for objective information on the effectiveness of educational software programs. Until we know how specific characteristics of software influence instructional effectiveness, educators can only guess at which programs will work best in their classrooms, software developers can only guess at which instructional and programming features make their software more effective as teaching tools, and reviewers of software will be limited to subjective judgments with questionable validity.

Classroom teachers can make an important contribution toward generating the required data base by conducting small evaluation studies of software that they are using. Although any single study is likely to have low statistical power and limited generality, the aggregate data acquired from many studies will be of great value in helping to identify specific factors that make educational software effective. □

Notice to Readers

Do you know of someone who would like to conduct an outcome evaluation of specific microcomputer software? We will provide standardized procedures for evaluation in different circumstances, compile the results of the evaluations, and make the results available to educators. For more information, please contact:

Dale E. Berger, Ph.D.
Software Evaluation Group
Psychology Department
The Claremont Graduate School
Claremont, CA 91711-6175

References

Campbell, D.T., and Fiske, D.W. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 1959, *56*, 81-105.

Davis, W.D. An Empirical Assessment of Selected Computer Software Purported to Raise SAT Scores Significantly When Utilized with Short-Term Computer-Assisted Instruction on the Microcomputer. Unpublished doctoral dissertation, The Florida State University, College of Education, 1985.

EPIE Institute. *TESS: The Educational Software Selector.* 1985 Edition. New York: EPIE Institute and Teachers College Press, 1985.

EPIE Institute. *TESS: The Educational Software Selector.* 1986 Computer On-line Service Available Through CompuServe. New York: EPIE Institute, 1986.

Kulik, J.A., Bangert, R.L., and Williams, G.W. Effects of Computer-Based Teaching on Secondary School Students. *Journal of Educational Psychology*, 1983, *75*, 19-26.

Kulik, J.A., Kulik, C.C., and Cohen, P.A. Effectiveness of Computer-Based College Teaching: A Meta Analysis of Findings. *Review of Educational Research*, 1980, *50*, 525-544.

New Releases At-A-Glance. *Electronic Learning* (All issues, October 1985 through June 1986).

Nunnally, J.C. *Psychometric Theory* (2nd ed.). New York: McGraw-Hill, 1978.

Otte, R.B. Courseware for the 80's. *T.H.E. Journal*, October 1984, 89-97.

Schiffman, S.S. Software Infusion: Using Computers to Enhance Instruction. Part Two: What Kind of Training Does Software Infusion Require? *Educational Technology*, February 1986, *26*(2), 9-15.

Thomas, D.B. The Effectiveness of CAI in Secondary Schools. *AEDS Journal*, 1979, *12*, 103-116.

Watkins, M.W., and Abram, S. Reading CAI with First Grade Students. *The Computing Teacher*, April 1985, 43-45.