The Value of Supplementing Panel Software Reviews with Field Observations

Ronald D. Owston
Herbert H. Wideman

Abstract: When purchasing software for classroom use, educators frequently have to rely on software evaluation reports in making their decisions. Unfortunately, most reports do not make clear the extent to which the software being reviewed has been field tested or whether it has been field tested at all. In this study, teacher panel reviews of software were compared to field test reports to determine the levels of agreement between the two evaluation types and what kinds of additional information can be obtained from field observations. The results suggest that field testing may: a) bring to light technical and design limitations that are not obvious to teacher reviewers; b) provide more accurate information on the ease of use of the software; c) suggest unique ways in which the software can be used in the classroom; and d) give a clearer indication of the suitability of software in meeting specialized student needs.

Educators interested in carrying out summative evaluations of microcomputer software face no shortage in evaluation models from which they can choose. A search of the literature on software evaluation can easily turn up over 50 different evaluation forms, checklists, or complete models, and there are as well perhaps hundreds of unpublished forms developed for local use. All of these procedures usually provide a series of questions or a set of rating scales to guide the teacher or other expert in assessing the content and instructional quality of software. They normally place little, if any, emphasis on gathering data by means of direct observation of students' use of a program and on reporting these data. This is true of even the most widely used evaluation methods; for example, the evaluation guidelines published by the National Council of Teachers of Mathematics recommend only that 'two or three' students be observed during one use of the program, with most of the data for the evaluation being collected in the course of the teacher's use of the package (NCTM, 1984). MicroSIFT's widely used evaluation procedures do not require evaluators to observe students working with the software to complete their checklist, although this is left as an option (ICCE, 1984). MicroSIFT had intended to engage in field testing of software as the

Ronald D. Owston is Associate Professor of Education and Director of the Centre for the Study of Computers in Education, Faculty of Education, York University, 4700 Keele St., North York, ON M3J 1P3. Herbert **H. Wideman** is Research Associate in the Centre for the Study of Computers in Education.

CJEC, VOL. 16, NO. 4, PAGES 295 - 308, ISSN 0710 4340

fourth stage in their own evaluation activities, but this was never done. The Educational Products Information Exchange does require its evaluators to field test software until a consensus is reached about its quality, but there is no reference to any field test findings in their published reports (EPIE, 1986).

As the EPIE case illustrates, even when the evaluation procedures themselves may specify that field testing bc carried out, often there is no clear indication from the evaluation report whether, or to what extent, it was actually undertaken. The reader is typically given summary ratings and evaluative comments. Seldom is there any indication of the number of evaluators used, the length of the evaluation observation period, student responses, the type of school and classroom, or the age/grade of the students involved in the field test.

Nearly all the major models of educational evaluation, from Tyler's objectives-based design to more recent formulations such as Guba and Lincoln's responsive model and Eisner's advocacy of educational connoisseurship, emphasize the importance of gathering data from program users (Eisner, 1979; Guba & Lincoln, 1981; Tyler, 1950). More recently, there has been increasing demand that courseware reviews meet these criteria (e.g., Muller, 1985; Pike, 1983; Ragsdale, 1982; Tovar & Barker, 1986).

Evaluation consumers need to know whether an evaluation report is based on field testing because the ultimate test of the value of a software package is its effectiveness in the classroom with users. Software producers certainly cannot be relied upon to provide this information; a recent survey of producers found that only half of the 125 respondents did any form of field testing, and just 12% of those included the results in their product documentation (Truett & Ho, 1986). Educators may have the expertise to assess the quality of a program's content and instructional design, but it seems unreasonable to assume that they will be able to judge with any precision how students will respond to it. There is practically no research available that investigates how well teacher reviews are able to predict the classroom effectiveness of software. Certainly, if a consumer reads an evaluation report that was not based upon classroom observation, he or she has no reliable way of knowing how effective the software would actually be. The results of one study that has compared students' and teachers' perceptions of a series of algebra programs designed to complement classroom instruction do not encourage reliance solely on teacher ratings. Students proved to be stronger critics of the software than the teachers, rating it lower on adequacy of pacing and on instructional and motivational value (Signer, 1983). Other research suggests that, even when well trained, teachers are often uncritical of software they evaluate (Preece & Jones, 1985). Elementary and middle school students, on the other hand, can frequently make 'mature and sophisticated' judgements about a program's quality (Smith & Keep, 1986).

Experimental and quasi-experimental research has recently been advocated as being the most rigorous means for judging a program's effects (e.g., Muller, 1985; Tovar Barker, 1986). While these forms of evaluation can be most effective in minimizing threats to the validity of some forms of outcome data, practical limitations on the resources available for software assessment often restrict the number of packages that can be evaluated in this manner. Alternatively, the use of trained teachers, working independently, to observe and record student reactions to software and to

conduct student interviews would greatly increase the number of field tests that could be undertaken. While there would be an inevitable decrease in the amount of information gleaned about program outcomes, the collection of dam on students' affective and cognitive responses to different aspects of a program's content, design, and technical quality should provide a significant body of information which would substantially enhance the quality of teacher software reviews.

THE STUDY

The present study was designed to provide an assessment of the incremental value of one form of qualitative field testing undertaken by teachers working independently. A number of programs previously evaluated by panels of educators were field tested by teachers in their classrooms. (Both evaluation procedures are discussed in some detail below.) Qualitative comparisons were made between the panel and field evaluations of each software package in order to determine the extent to which the two kinds of data diverge and to assess the value of the contribution of the additional data obtained from the field test to the overall evaluation of the software. The data analysis had two parts. First, each comment and scale rating in the panel evaluations was coded as agreeing, disagreeing, or being supplemental to the comments in the matching field test report. In addition, field test remarks that contained information supplemental to that in the panel evaluations were coded as such. Coding was completed by each author independently; inter-rater agreement was found to be over 90%. The few discrepancies were resolved through discussion. In the second phase of the analysis, comparisons were made between the different pairs of evaluations using the constant comparative method (Glaser & Strauss, 1967). Substantive theories were developed about the overall trends in agreement, disagreement, and supplementation between the two forms of evaluation. These served as the basis for assessing the incremental utility of the field tests.

Thirty-six commercially available software packages were studied covering primary and junior level mathematics, language arts, reading, and general problem-solving skills. These included a variety of types of software ranging from drill and practice to more open-ended problem-solving packages.

Evaluation Procedures

A two-stage summative software evaluation model developed at York University Faculty of Education makes clear the role of field testing in software assessment. The model separates the evaluation process into two distinct phases teacher review and field testing. Each phase has its own unique procedures, yet both yield evaluative data in the same categories.

Panel Evaluation. The first stage of the model is called panel evaluation (Owston, 1987). At this stage, groups of three teachers develop a consensus on what the rating of a software package should be using the York Educational Software Evaluation Scales (YESES). YESES is a set of four criterion-based scales that provide a more global or holistic assessment of software than is possible with the more commonly-used checklist approaches. It is more holistic in the sense that evaluators judge the overall quality of

the software using only four scales, rather than responding to a lengthy series of questions typically found in checklists. The four scales of YESES are pedagogical content, instructional presentation, documentation, and technical adequacy. The pedagogical content scale refers to the knowledge and skills that the software purports to teach, including their organization, accuracy, and appropriateness; the instructional presentation scale is concerned with the manner in which software takes advantage of the unique features of the microcomputer in presenting the content; the documentation scale refers to the supporting materials and instructions, available both in print and on screen, that accompany the software and explain its use; and the technical adequacy scale is concerned with the quality of the software design with respect to user inputs, software outputs, and system errors. For each of these dimensions, there is a four-point scale that describes the general characteristics of software that would be rated at that level. Software rated level 4 is labelled 'exemplary', level 3 'desirable', level 2 'minimally acceptable', and level 1 'deficient'. Figure 1 provides an illustration of the pedagogical content scale of YESES.

Figure 1. Pedagogical Scale of YESES.

DEFINITION

Content refers to the knowledge and skills the software purports to teach — the organization, accuracy, and appropriateness of the material. Content organization refers to the sequencing of the knowledge and skills within the lesson or lessons, the breadth or scope of the skills and knowledge, and the depth of instruction or amount of practice given to a topic. Accuracy is concerned with the truthfulnessof the knowledge and skills presented. Appropriateness deals with the suitability of the content for the intended user, including such factors as readability, the relationship between the complexity of the content and the intended user's ability to master it, and the educational value of the content (i.e., whether the time spent learning the content is justified because of its inherent value). If one or all of these elements — organization, accuracy, and appropriateness — are weak, the content may be judged less than exemplary.

Level 4: Exemplary Content

Level content is superior in its organization, accuracy, and appropriateness. The content organization is such that the scope of the knowledge and skills is congruent with the user's ability to master them, the sequencing is logical and follows good pedagogical practice (e.g., less abstract ideas are presented before more abstract ideas), and the depth of instruction is sufficient to give the user adequate practice before proceeding to the next topic. The accuracy of Level 4 content is extremely high. Furthermore, the content at this level is very readable, well-matched to the intended user's ability to master it, and has high educational value.

Continued on next page.

Figure 1, continued. Pedagogical Scale of YESES.

Level 3: Desirable Content

The organization, accuracy, and/or appropriateness of Level 3 content is not quite as favorable as that of Level 4 due to relatively minor weaknesses. The organization may be weak because the content scope does not quite match the user's ability to master it; the sequencing may be illogical or not in keeping with accepted pedagogical practice; the intensity of instruction may be either slightly more or less than necessary, requiring the user to complete too many or too few exercises; and the user may not receive sufficient practice with the material before moving on to the next topic. Problems with accuracy might consist of questionable (but not incorrect) facts or applications of concepts. Level 3 content may also present some vocabulary or sentence structures that give intended users difficulty. Its material may be too complex or too easy for the intended user to digest, and some aspects of the content may be of questionable educational value. However, all flaws in Level 3 content are slight.

Level 2: Minimally Acceptable Content

Level 2 content is weak in either one area or a combination of the areas of organization, accuracy, or appropriateness. The deficiency, however, is not serious enough to prevent the use of the software, if no other better software is available, and if the instructor is able to rectify the deficiency. In its organization Level 2 software may present too much material; it may be poorly arranged in sequence or not consistent with good educational practice; its instructional depth may be exaggerated or insufficient. Accuracy problems encountered with Level 2 content include incorrect minor facts or applications of concepts. At this level vocabulary and content structure may be too difficult for the intended user, the knowledge and skills too difficult to master (or too easy), or the educational value of the overall content questionable.

Level 1: Deficient Content

Content at Level 1 is sufficiently substandard to call into question the use of the software, regardless of the strengths of its other characteristics. Organizational problems may include weak, illogical sequencing, and content scope and/or depth of instruction poorly matched with the user's ability. This Level of content may also contain factual inaccuracies or incorrect applications of concepts. The content reading level may be inappropriately matched with the user's ability, the knowledge and skills presented eithertoo complex or simple, the topicscovered of dubious educational value.

The design of YESES was influenced by developments in three areas. The first is the field of the assessment of writing and, in particular, the analytical method of scoring writing (Diederich, 1974). In this field, the assumption is that there are several identifiable underlying traits of writing, all of which, in any context, are considered important, upon which the writing can be judged. Scales are constructed to measure each of these key traits, with each scale point being explicitly defined to describe writing characteristics of that level. The second field is criterion-referenced testing

(Popham, 1978). Here the belief is that more meaningful assessments of achievement, for example, can be attained by determining the extent to which specific domains of knowledge have been mastered, rather than by basing the assessment on achievement relative to others taking the test. Assessment of second language oral proficiency is the third area from which the rationale for YESES is drawn. Specifically, those techniques that enable teachers, through loosely structured interviews, to rate a student's overall proficiency according to pre-defined criteria were examined (e.g., NBDE, 1974). The ideas of these three fields were influential in the design of YESES in that: a) four key characteristics of software were identified; b) criteria against which software would be judged were specified before evaluations were made; and c) evaluators have to make global assessments in each of the scale areas about the overall quality of the software.

In panel evaluation, the process is one of becoming thoroughly familiar with the software and then determining which level of YESES best characterizes the software for each of the four dimensions. The final step requires evaluators to write short evaluative comments, mentioning any unique features of the software and its possible applications, strengths, and weaknesses. When both the panel ratings and the written notes are combined, the reader obtains an overall impression of the quality of the software from the panel's perspective.

Evaluators were given a one-day training session on the use of the instrument. The training session consisted of an in-depth introduction to the rationale, design, and interpretation of YESES. This was followed by a 'hands-on' experience in which individuals evaluated a software package that has been rated previously by the original calibration group involved in the development of YESES. A group discussion was then held during which time the original calibration group's ratings of the software were revealed and evaluators were given the opportunity to raise questions and seek clarifications. Evaluators then had the opportunity to do at least one more practice evaluation.

The inter-rater reliability and validity of the panel evaluation process have been found to be reasonably high (Owston, 1985, Owston & Dudley-Marling, 1986). When a software package is rated by the same panel on two different occasions, or by different panels, the ratings are, without exception, found to differ by no more than one or two points on one, two or three of the four evaluation scales. This level of reliability is realized because of the training evaluators receive in using and interpreting YESES and because of the use of explicit criteria for the scales of YESES. When a sample of panel evaluations were compared to evaluations of the same products done by EPIE, there was agreement or partial agreement on the overall value of the products in 71% of the cases. Further support for the validity of the YESES procedures can be found by observing the high level of agreement between software packages that are 'recommended' by Alberta Education (1986) and those that receive 'exemplary' ratings on all four scales of YESES.

Field Testing. Field testing, the second of the two stages in the York evaluation model, requires that software be tried out in a classroom for four to six weeks (Dudley-Marling, Owston & Searle, 1986). Common guidelines for conducting field tests were given to teachers and a group discussion was held on their use. The guidelines call for the teacher to observe children using the software and informally record their reactions. These notes are later used by the teacher for organizing a discussion with students

about the software. Both the teacher observational data and student data are then grouped into the categories of YESES that most appropriately describe them. At this point a narrative report of two to three pages is written. In addition to the observations, the teacher reports on the dates when the evaluation was carried out, the hardware used, and the curricular context and instructional setting in which the evaluation took place. Only teachers who had had experience in conducting panel evaluations were used for field testing software. Furthermore, teachers did not field test the same products that they evaluated in panels.

In brief, the field test report describes how one teacher used a given software package in a particular setting and what results were obtained. The extent to which the results would be applicable to other settings will depend upon the similarity of the two settings. Thus the field test should be viewed as information to supplement the panel evaluation, not as a summative evaluation in its own right.

RESULTS

The comparison of the panel evaluations with the corresponding field tests proved illuminating. In 19 of the 36 pairs of evaluations that we have studied, the two forms of software review have been in general agreement about the quality of the software tested. In 10 cases, the two evaluations offered strongly divergent assessments of the quality of the software. And in the remaining seven instances, the panel and field evaluations concurred on certain aspects of the software's quality but disagreed on others.

For the programs about which there was general agreement, those given 'desirable' or 'exemplary' ratings for content and instructional presentation in the panel evaluations were found by the field testers to be effective learning aids that held students' interest and made good use of the computer's capabilities. The packages considered inadequate or barely acceptable by the panels were those that the field testing indicated had little educational utility because of inappropriate content or poor design. The evaluations and field tests were usually also in agreement about several other aspects of the software, such as the quality of the screen displays and the documentation, the ease of program operation, and the adequacy of the feedback and on-line help.

The evaluations of ***The Puzzler*** program, shown in Figure 2 (see ***next page),*** provide an illustration of the level of agreement between the panel and field findings typical of the majority of cases. The panel evaluators rated the content, instruction, and technical quality of the program as 'desirable', and the documentation as 'exemplary'. They noted in their comments that the novelty of the material should stimulate interest, and that the program invited divergent thinking and shared reading experience.

The students' experiences with the software, as reported by the field tester, seemed to bear out the panel's evaluation. The teacher noted that the students maintained a high level of interest in the program and completed all of the relevant tasks. The students cooperated with each other in problem-solving, interacting frequently as they analyzed choices and made predictions. The documentation, which the panel had considered comprehensive and exemplary, proved extremely helpful to the teacher in integrating the program with the curriculum.

Figure 2. Sample Panel Evaluation and Field Test Report.

YORK PANEL EVALUATION RESULTS

DATE OF EVALUATION (YY/MM/DD): 85/03/01
EQUIPMENT USED: Apple IIe
EVALUATORS: SE., L.H., J.F.
TITLE: The Puzzler
RATINGS (4-exemplary, 3-desirable, 2-minimally acceptable, 1 -deficient)

CONTENT	3	DOCUMENTATION	4
INSTRUCTION	3	TECHNICAL	3

Comments:

The content of this program is based upon a sound theoretical framework, reflecting contemporary approaches to the teaching of reading. The material offers a supplement, although not a substitute for the regular reading materials. The effects of novelty could serve to stimulate interest particularly for the less able student. The open-ended nature of the material invites divergent thinking and permits shared reading experience. The documentation offers a comprehensive package which clearly integrates the computer software into the context of the total reading program.

YORK FIELD TEST RESULTS

DATE OF EVALUATION: February – March 1986
EQUIPMENT USED: Apple IIe
EVALUATOR: P.T.
TITLE: The Puzzler

Comments

The *Puzzler* was field-tested over a five week period in a Grade 4 class of 28 students. With two computers in use in the far corners of the classroom, the children worked in pairs on a rotation schedule. Each pair of students had access to the program for 25 minutes per day. The majority of students had two years previous experience on microcomputers, although this was the first reading program they had encountered. The *Puzzler* was used as a challenge and supplement to the regular reading curriculum.

The five stories, of increasing length and difficulty, generated a high interest level in these nine and ten year-old children. Three of the five selections were written in the first person, and two were animal stories. Students commented that the stories were imaginative and challenging. The children were able to read to the end of a particular story and complete the predicting and confirming tasks without losing interest.

The program promoted a spirit of cooperation rather than competition between the students. There was a good deal of interactive language as they

Continued on next page.

Figure 2, continued. Sample Panel Evaluation and Field Test Report.

made predictions and analysed their choices. However, it was observed that towards the end of the test period, once the children had read the five selections several times and were familiar with the stories, they were not eager to the program. Perhaps there could have been more stories on the disk.

In terms of technical adequacy, the key commands were satisfactory and straight forward, although not elaborate (Arrow keys, S for Story, P for Prediction, etc.). However, the Escape option did not always work, leaving the reader with little flexibility in skipping ahead or backwards to other selections. Similarly, access to the menu was limited, and students wished they had more control over the software than simple page turning.

The teaching strategies outlined in the written documentation proved to be extremely helpful in introducing The Puzzlerto the class. Using an overhead projector, two sample stories "Petoskeys" and "The were read by the group. Predictions were made based on contextual clues, with students being given an opportunity to discuss and modify their choices. These two sample lessons prepared the students to read carefully and not be anxious about finding a 'correct' answer.

As an extension to this reading program, a group of seven students were motivated to compose their own open-ended stories. These were written individually and in pairs, both in longhand and on a word processor. The stories were read and discussed by their peers. This language arts component of the reading program showed that the students had understood the principles underlying *The Puzzler* and were able to assimilate them in a creative way.

As the documentation states, the children's transfer of predicting and confirming strategies to their daily reading is proof of the value of the experiences they received from *The Puzzler.* This transfer of skills was observed in daily reading assignments where students stayed on task a little longer than expected, despite difficult vocabulary levels. Students' comments were that The Puzzler "helps you learn," "gets you thinking hard," "keeps you reading until you figure it out," and "is fun once you know the story."

As well as confirming (or occasionally disputing) the panel's analysis, the field

field report highlighted some minor limitations in the design of the program that limited users' access to the program's menu. It noted that the addition of more stories to the disk could extend the utility of the package. In addition, the report offered a useful illustration of the ways in which the program activities could be effectively transferred to off-computer tasks in ways that would reinforce and extend the students' new learnings. And finally, it reported some evidence for the students' transfer of newly-mastered skills to other domains.

The panel and field evaluations were as likely to agree that a program was of high quality as they were to agree that a package was mediocre or inadequate. There was no evidence of a ceiling or floor effect in the rating levels for either type of evaluation. However, when the two forms of evaluation differed in their assessment of a package,

it was usually the panel report that was more critical of the software.

There were ten packages over which there were major differences between the two evaluations about the software's quality. In each case, it was the evaluation panel that felt the package to be inadequate, while the field test report rated the package highly. Our analysis of the qualitative data indicated several reasons for this phenomenon.

First, and most importantly, there were several instances in which a teacher using the package in the field test modified or structured its use in a manner not suggested in the program's documentation, with the result being that the software was used to much better effect. These new patterns of use had not been anticipated by the panel, which had worked under the assumption that the programs would be employed in a straightforward, 'plain-vanilla' fashion, and had evaluated them on that basis. The evaluation and field testing of *Tales of Adventure* exemplify this pattern of discrepancy. (Students using this story program choose which of several directions they wish the plot to take at different points in a tale.) The panel evaluators rated this program as 'deficient' in all categories, and commented that the storylines were simplistic and unrealistic and that the software generally lacked usable content. However, the teacher was able to overcome the program's apparent limitations by integrating it into a structured lesson, which included a number of program-related activities:

> We went through the first few screens as a group to help children get started. Pupils were then divided into groups of four. Each group kept track of the different paths they took on index cards. After each group had 'experienced' the program at least once, the class as a whole was introduced to flowcharts. They learned the three types of symbols used in flowcharts. Next each group was asked to flowchart parts of the stories on the disk on chart paper. The next step was for each group to create their own adventure. This was first done on chart paper. The flowchart was then developed into pages that eventually were published as books.

The field tester found the outcomes of this structured approach to use to be favourable:

> ***Tales of Adventure*** is an attractive program. It can be a springboard for stimulating conversation, language development and social interaction. Children in higher grades effectively tutored those in lower grades in its use. Language was used to direct, report, predict, hypothesize and imagine.

A similar pattern can be seen in the evaluations of the program *All About Dinosaurs*. The panel considered this program deficient in content and instructional design, noting that-with the exception of a branching story section -the material was presented in a tutorial-quiz format more appropriate to a book. While the field evaluator agreed that the tutorial material was generally inadequate and made little use of the computer's features, she was able to make effective use of the branching story part of the program, but only by structuring the activity carefully: "Timely intervention by the teacher in encouraging the prediction and discussion of possible outcomes for each situation led to higher order thinking and problem-solving by the students. If this

intervention did not occur, student choices were often made randomly with little discussion or verbal input."

In a few instances, a program that had been given a low rating by the panel was used with apparent effectiveness by students with special needs. For example, *Alphabet Zoo,* a program designed to enhance letter recognition and improve spelling, was found to be beneficial for children having specific difficulties:

> This program proved to be especially effective when a more able student was paired with one experiencing difficulty with letter recognition. -The program provided opportunities for oral discussion as the children attempted to guess what picture was being drawn by the computer . . . If students are unsure of the letter name they are assisted by their partners.

In addition, an autistic child found the program to be highly interesting. He was able to identify many pictures independently and repeat the letters as they were articulated by a learning aide.

Panelists' disagreement with the underlying educational philosophy of a program (as inferred from the program's design) was another reason for the divergence in ratings between the panel and field reports. It resulted in the panels downgrading three packages. *The Game Show* was rated as 'deficient' in content and instruction by the panelists, who stated that "the program is very limited in its view of knowledge, and focuses on one-word answers without encouraging much student thought." The teacher who conducted the field testing had a different view: "This program can be used effectively to build vocabulary, reinforce spelling skills, and provide opportunities for verbalization, cooperation, and teamwork." It is clear that she had a different perspective on what pedagogical goals the program was meant to serve, and in her view it met *those goals* well. *Comprehension Power,* given a minimally acceptable rating in all four categories by panel evaluators, was also downgraded for its theoretical inadequacies: "A simplistic view of comprehension underlies this program. Students focus on selecting a right answer rather than developing personal understanding." The teacher conducting the field test, however, noted that the students felt the program offered a far superior alternative to paper and pencil testing of comprehension. The students were challenged by the program, and the teacher felt no conflict between her own educational style and that offered in the program; in fact, she indicated that it integrated very well into the curriculum.

Four other programs considered effective by field testers were heavily criticized by the panels for their overly simple design. In their comments the panelists noted a lack of appropriate feedback to the child, the simplistic forms of response allowed, and the limited range of the programs. However, the field trials suggested that even with these flaws the software still interested students, and could effectively facilitate learning and problem solving.

Surprisingly or not, depending upon one's perspective, panelists occasionally experienced difficulty in using a piece of software that young children found relatively easy to use, and this was another source of divergence in program ratings. In reviewing *Turtle Power,* a Logo subset for young children, the panelists noted several operational

limitations (such as the need to save and load program subroutines separately) that were considered likely to inhibit effective use of the program. However, the field test report indicated that children could use the package efficiently, and that no technical problems were noted. Similarly, the panel evaluating *Gertrude's Secrets* commented that they had some difficulty in learning to operate the program and that children may require extensive support from the teacher in order to be able to do so. The field tester found, however, that even pre-readers were able to learn how to solve the puzzles simply by looking into the sample puzzle rooms.

In the case of a few other packages, panelists were more critical than field testers of the quality of the program documentation and/or the legibility of screen text and graphics. (It should be noted, though, that for several other programs, panelists were able to accurately anticipate the degree of difficulty users would have with software and documentation.)

For seven pairs of evaluations, what we have termed 'partial agreement' was found between the two forms of assessment. This level of agreement was defined as occurring when one of the evaluations considered the software under review to be mediocre (because of either several minor or one major weakness) while the other reported it to be either very desirable or not worth using. In three of these instances, field evaluators were more critical of the software than the panelists. The field testers cited technical problems with program operation, or noted a lack of student interest in the program.

As noted previously, the field tests often provided information that significantly supplemented that obtained from the panel evaluations. In several instances the field reports indicated that programs were used effectively as a means to promote cooperative problem-solving and the development of communicative and social skills. These outcomes were not always anticipated by the panel evaluators.

The reports usually provided other supplementary information of significance. The levels of interest generated by a program and whether these were sustained over time would be discussed. Field reports often indicated that students enjoyed using a program more when they had a greater choice in their strategies and more control over the computer. And they sometimes highlighted ways of integrating the software into the curriculum that did not occur to the panelists. On a few occasions the field evaluations broke new ground by pointing to parts of the program where better remediation or on-line help were necessary. And in two trials, the field testing indicated that the software might be more appropriate for a different age range.

CONCLUSION

The results of the comparative analysis suggest that, where possible, field tests of software should be undertaken to supplement other forms of evaluation. Field tests, by providing additional data on which to base evaluative judgement, can serve to increase the utility of the evaluation report in a number of ways. Classroom observation can bring to light technical and design limitations in software that may not be apparent to adult users. Conversely, operational difficulties foreseen by adult evaluators may prove to be minimal for certain student groups. A broader range of educational perspectives

and philosophies can be brought to bear in the assessment of a program's value, including the students'. Field reports, by presenting the educational contexts in which software is used, can illustrate creative and effective strategies for program use and for its integration into the curriculum in ways that might not have been anticipated by evaluators, resulting in a greater appreciation of the program's potential. Field tests may also give a clearer indication of the suitability of a program for meeting certain specialized student needs.

This is not to argue that software evaluations that lack a field component have no value, however. The overall level of agreement found between the panel and field tests in the present study suggest that systematically conducted panel evaluations will usually be able to assess a program's quality with reasonable accuracy. Our findings do indicate that evaluators need to be sensitive to a wide range of contexts and strategies for a program's possible use if they are to do it justice in their assessment. But the ease with which simple field evaluations (of the type employed in the present study) can be implemented, and the value of the information to be gained, argues strongly for undertaking field testing as a part of software evaluation far more frequently than is presently the case.

REFERENCES

Alberta Education. (1986). *Computer courseware evaluations: June, 1985 to March, 1986.* Edmonton, AB: author.

Dudley-Marling, C., Owston, R. D., & Searle, D. (1986, October). *Guidelines for the field testing of microcomputer software* (York/IBM Cooperative Project Document No. 5). North York, ON: York University Faculty of Education.

Diederich, B. (1974). *Measuring growth in English.* Urbana, IL: National Council of Teachers of English.

Eisner, E. W. (1979). *The educational imagination.* New York: Macmillan.

EPIE Institute. (1986). *The educational software selector.* New York: Columbia Teachers College Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory.* Chicago: Aldine.

Guba, E. G., & Lincoln, *Y. S.* (1981). *Effective evaluation.* San Francisco: Jossey-Bass.

International Council for Computers in Education. (1984). *Evaluator's guide for microcomputer-based instructional packages.* Eugene, OR: University of Oregon.

Muller, E. W. (1985). Application of experimental and quasi-experimental research designs to educational software evaluation. *Educational Technology, 25*(10), 27-31.

National Council of Teachers of Mathematics. (1984). *Guidelines for evaluating computerized instructional materials* (rev. ed.). Reston, VA: author.

New Brunswick Department of Education. (1974). *Manual for interviewers of French.* Princeton, NJ: The Educational Testing Service.

Owston, R. D. (1985, December). *Software evaluation using YESES.* Paper presented at the annual conference of the Ontario Educational Research Association, Toronto.

Owston, R. D. (1987). *Software evaluation: A criterion-based approach.* Scarborough, ON: Prentice-Hall.

Owston, R. D., & Dudley-Marling, C. (1986, April). *A criterion based approach to software evaluation.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Pike, R. (1983). *Evaluating the effectiveness of lessonware prototypes: Some guidelines for use in the development of educational software.* Toronto: Ontario Ministry of Education.

Popham, W. J. (1978). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Preece, J., & Jones, A. (1985). Training teachers to select educational software: Results of a formative evaluation of an Open University pack. *British Journal of Educational Technology, 16*(l), 9-20.

Ragsdale, R. G. (1982). *Evaluation of microcomputer courseware.* Toronto: OISE Press.

Signer, B. (1983). How do teacher and student evaluations of CAI compare? *The Computing Teacher, 11*(2), 34-36.

Smith, D., & Keep, R. (1986). Children's opinions of educational software. *Educational Research, 28*(2), 83-88.

Tovar, M., & Barker, N. (1986). Field test evaluation of educational software: A description of one approach. *Canadian Journal of Educational Communication, 15*(3), 177-190.

Truett, C., & Ho, C. (1986). Is educational software fieldtested? *The Computing Teacher, 14*(2), 24-25.

Tyler, R. W. (1950). *Basic principles of curriculum and instruction.* Chicago: University of Chicago Press.