DOCUMENT RESUME

ED 324 337 TM 015 553

AUTHOR Barrett, Andrew J.; And Others
TITLE Objectively Determining the Educational Potential of Computer and Video-Based Courseware; or, Producing Reliable Evaluations Despite the Dog and Pony Show.
PUB DATE Mar 90
NOTE 18p.; Paper presented at the the International Conference on Technology and Education (7th, Brussels, Belgium, March 20-22, 1990).
PUB TYPE Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Assisted Instruction; Computer Software; *Computer Software Evaluation; Content Validity; *Courseware; Educational Assessment; Elementary Secondary Education; *Evaluation Methods; Evaluators; Higher Education; *Instructional Effectiveness; Models; *Reliability; Training Methods; Validity; *Videotape Recordings
IDENTIFIERS Center Interactive Technology Applications Res; Objective Analysis; University of South Florida

ABSTRACT

The Center for Interactive Technology, Applications, and Research at the College of Engineering of the University of South Florida (Tampa) has developed objective and descriptive evaluation models to assist in determining the educational potential of computer and video courseware. The computer-based courseware evaluation model and the video-based courseware evaluation model provide information on over 300 items, 150 of which pertain directly to instruction (e.g., defining objectives, instructor skills, lesson characteristics, questioning and tutorial techniques, learner interactions, resource scope, and content assessment). Descriptive data were obtained in one of three manners: (1) a yes/no indication of whether or not a characteristic is found; (2) a 4-point ranking indicating the necessity of a characteristic to use the package effectively; and (3) a 4-point Likert-type scale revealing the extent to which a given characteristic is used by the courseware. Content validity was investigated by four experts evaluating 10 computer-assisted learning packages and five experts evaluating seven video-based packages. Reliability was investigated as each evaluator produced synthesized quality ratings for each course. Results suggest that both models provide fairly close agreement estimates and fairly reliable evaluations. An overview of 213 computer-assisted learning packages and 248 video-based training packages indicated that many developers have failed to take advantage of their inherent capabilities, with particular shortcomings in: management system capacities; the instructor's control of lessons; and the student's flexibility in moving through lessons. Nine figures and two tables contain study data. (SLD)

Objectively Determining the Educational Potential of Computer and Video-Based Courseware; or, Producing Reliable Evaluations Despite the Dog and Pony Show

by:

Andrew. J. Barrett, Ed.D.
Theodore Micceri, Ph.D.
William H.Pritchard, Jr.

Center for Interactive Technologies, Applications and Research (CITAR)
Assistant Dean's Office
College of Engineering
USF
Tampa, FL 33620, USA
(813) 974-3783

Paper presented at the Seventh International Conference on Technology and Education, Brussels, Belgium, March 20-22, 1990

ABSTRACT

In today's complex market for technology-mediated instruction, flashy presentations frequently prove the most important purchasing element while instructional design and content take a back seat to form. Harried consumers faced with a plethora of information and bombarded by sophisticated visual effects must base purchasing decisions upon either a vendor's presentation or some published evaluator rating that is almost certain to be biased by irrelevant factors. The Center for Interactive Technology, Applications and Research (CITAR) has developed an objective and descriptive evaluation model to ameliorate this situation.

INTRODUCTION

At present, a massive amount of educational software is available to the consumer. Frequently, much of this software is highly priced, and all too often, of poor quality. Even though a growing number of vendors allow evaluation copies before purchase, the selection process is time-consuming and labor-intensive. Many publications and professional organizations review technology-mediated courseware for their clientele, however, most reviews are highly subjective and cannot be used to compare or rank packages of similar content against each other. Purchasers require objective courseware evaluations in order to

make informed decisions. To address this problem, CITAR began a project to develop descriptive courseware evaluation models. During the past two years the CITAR Computer-Based Courseware Evaluation Model (CCCEM) and Video-Based Courseware Evaluation Model (CVCEM) have been used to analyze numerous educational and training courses. Each of these evaluation models provides descriptive information on over 300 different items, some 150 of which pertain directly to instruction and deal with topics such as: defining objectives, instructor skills, lesson characteristics, questioning and tutorial techniques, learner interactions, resource scope, content assessment, etc. The purpose of this development was to provide an objective and comprehensive evaluation model for the assessment of technology-mediated instruction. Since any such measure must include a human evaluator, the most important task facing developers was to assure that evaluations of the same courseware by different evaluators produced essentially the same results. A second problem concerned an evaluator's ability to synthesize such vast amounts of information into overall ratings that allow for comparisons across similar courseware packages.

DEVELOPMENT OF EVALUATION MODELS

To avoid the problems of positive bias that universally associate with qualitative ratings, a purely descriptive model was developed requiring evaluators only to determine the existence or non-existence and frequency of occurrence for specific courseware characteristics. This separates the process of measurement from that of evaluation, thereby reducing error. Historically, this has proven the most efficient psychometric method for creating objective performance evaluation measures (Peterson, Micceri & Smith, 1985). Additionally, to enhance evaluator accuracy, the models break complex concepts such as instruction, management and user interface into component pieces small enough to be objectively defined. In this way, both the existence and prevalence of various content and technological aspects of computer courseware may be compared across similar courses.

To ensure validity, evaluations are conducted by graduate students from the University of South Florida each of whom is a content expert on the relevant topic. Every evaluator receives a minimum of 20 hours of training on the model and is checked for accuracy on representative packages before conducting evaluations.

Descriptive data are obtained in one of three ways: (1) Yes/No — indicating whether the courseware contains a characteristic; (2) a four point ranking — indicating the necessity of a characteristic to effectively use the package (RONSI: Required, Optional, Not Supported, or Indeterminate); and (3) a four point Likert-type scale — revealing the extent to

which a given characteristic is used by the courseware (ESON: Extensively [90% or more of the time]; Significantly [50-89%]; Occasionally [11-49%]; and Negligibly [1-10%]).

As noted, each of these models reports descriptive information on over 250 items. For a purchaser or evaluator, the synthesis of so many variables into a reasonable description is impossible and tends to produce inconsistent overall ratings that are more likely to indicate personal biases than "true" differences in software. A solution to this problem was sought for the purchaser, and perhaps found, when experts from a variety of interest groups determined that all individual items fall into one of four basic conceptual areas: Physical (19 CCCEM, 22 CVCEM items), Presentation (31, 99 items), Instruction (115, 125 items) and Management (51, 61 items). Within each of these topical areas, it proved feasible to develop scores capable of differentiating among courseware and amenable to rigorous scientific tests. Therefore, in addition to purely descriptive information at the item level, four major qualitative scores were created from weighted CCCEM items to service the needs of varying audiences and corresponding with the conceptual areas identified above. These scores were then combined into an overall score designated: courseware EFFICACY, which is heavily weighted on instructional and management components.

OUTLINE OF STUDIES AND METHODS

To evaluate content validity, a panel of experts having over 50 years experience in technology-mediated instruction was assembled. Following this, two G-studies were conducted, one for each model, to determine the level of agreement among raters on individual items, and the reliability of qualitative scores and subscores. Study 1 involved ten Computer Assisted Learning (CAL) packages, each of which was rated by four evaluators. Study 2 compared seven video-based instruction packages each of which was assessed by five evaluators.

Reliability may be viewed as the consistency with which an instrument differentiates among a group of targets (courses). Put simply, the consistency with which the same relative rankings are assigned to a specific group of targets by different raters. A major source of error for observation instruments is disagreement among raters, thus this issue was considered separately from the overall reliability question to isolate its influence. Interrater agreement may be defined as the extent to which two or more observers, working independently, agree on which phenomena occur to what degree in the target of interest. In this study, agreement represents the mean percentage across all items (i), all rater pairs (kr) and all subjects (j) for each score, with score level agreement defined as:

$$1.0 \quad R_a = \frac{x_i - x_{(i+1)}}{x_{hi} - x_{lo}}$$

where:

x_{hi}=highest obtained value in sample

x_{lo}=lowest obtained value in sample

This technique, although an excellent measure of the absolute magnitude one type of error (observer disagreement), provides no information about an instrument's ability to differentiate among targets, therefore, extremely high interrater agreement percentages may associate with very low reliabilities and vice versa. Although numerous techniques have been suggested for the investigation of observation instrument reliability, consensus appears to support the use of generalizability theory and the intraclass correlation coefficient (ICC) based on Analysis of variance (Mitchell, 1979; Shrout and Fleiss, 1978). The ICC coefficient is computed using an ANOVA table: with between target variance (BMS) treated as the "true variability" and within cell variance (WMS - combining instrument, time and rater error) treated as error:

$$2.0 \quad R_{ICC} = \frac{BMS - WMS}{BMS + (k-1)WMS}$$

where:

BMS = between targets mean square

WMS = error or within targets mean square

k = number of raters/judges

Additionally, test-retest reliability estimated were computed for all scores using Pearson's *r*.

In order to determine whether a gain in reliability (consistency) occurs by using relatively objective criteria such as those in the CVCEM model over the type of subjective overall evaluations by raters that are typically used to evaluate courseware, each evaluator also produced synthesized quality ratings for each course in both studies. These ratings are designated "perceptual" scores in following sections Each perceptual rating corresponded to one of the five major CVCEM scores: Instruction, Management, Presentation, Physical and total Efficacy. Ratings were based on a scale from zero to 10, where zero represents non-existent or useless and 10 represents optimal. Agreement percentages using formula 1.0 and

ICC estimates using formula 2.0 were then computed for these perceptions and compared with results produced by the same evaluators using the objective CITAR evaluation models.

Following these G studies, evaluations were conducted on 213 CAL packages and 248 video-based training packages. Some selected results from the G studies and certain descriptive information obtained from the numerous courseware evaluations is reported below.

RESULTS AND DISCUSSION

G-Study Results

For the most important and heavily weighted scores, Study 1 (CCCEM) produced agreements averaging about .83 (Table 1) and ICC reliabilities around .70 (Table 2). Study 2 (CVCEM) produced agreements averaging about .85 (Table 1) and also had ICC reliabilities around .70 for most scores (Table 2). Test-retest reliability results averaged approximately 10 percentage points higher than ICC estimates for most scores. Complete results and detail may be found in Micceri (1988 and 1989).

Table 1 compares agreement estimates computed separately for perceptual and evaluation model scores produced in the two studies. It is clear that except for the Physical subscore, the agreement estimates are fairly close, although the CITAR evaluation models consistently hold an advantage of between 10 and 15 percentage points.

TABLE 1
Agreement Estimates Comparing Perceptual Ratings and CITAR Evaluation Model Scores

	Study 1		Study 2	
Score	Perceptions	CCCEM	Perceptions	CVCEM
Efficacy	.69	.83	.76	.88
Instruction	.61	.80	.74	.84
Management	.64	.87	.92	.94
Presentation	.65	.79	.73	.85
Physical	.69	.91	.66	.93

Table 2 compares ICC reliability estimates for these two different approaches, and it is here that important differences appear. In Study 1, perceptual scores produced ICC estimates that range between .08 and .23. In Study 2, with the exception of the Management score, these ICC estimates range between .17 and .38. The management subscore was more reliable

because only one course had a management system. With the exception of the Presentation score in study 1 (.17) no CITAR evaluation model ICC estimates are below .55.

Table 2

ICC Estimates Comparing Perceptual Ratings and CITAR Evaluation Model Scores

	Study 1		Study 2	
Score	Perceptions	CCCEM	Perceptions	CVCEM
Efficacy	.18	.65	.35	.79
Instruction	.23	.66	.38	.58
Management	.10	.67	.88	.95
Presentation	.08	.17	.22	.59
Physical	.11	.55	.17	.71

These results suggest that both the CCCEM and CVCEM provide fairly reliable evaluations of courseware in both an absolute and relative sense. It is only for the Presentation score in the CCCEM that reliabilities are unacceptably low. Interestingly, this is the area to which many commercially available evaluations pay the greatest attention.

AN OVERVIEW OF EVALUATED COURSEWARE

CAL Courseware: CCCEM

The data on hand suggest that most courseware producers make little use of the rich interactive environment in which they work. For example, many CAL packages claiming interactivity are not truly interactive but use concurrent applications. Fewer than 5% of the CAL packages evaluated use interfaces other than the keyboard (note that these courses are primarily directed toward the IBM PC environment). Fewer than 10% generate reports for the instructor/manager regarding student performance. Thus, from a teacher/manager perspective, the current courseware lacks essential components. Planning also appears a weak area, with fewer than 25% identifying expected entry or exit proficiencies and only 8.3% defining measurable outcomes. Figure 1 shows cumulative percentages relative to optimal scores on the four major CCCEM scores. Maximum possible scores for the various scales represent what to the best current knowledge are relatively OPTIMAL learning environments. Scores are reported as a proportion of this optimal possible score. Other than for Physical characteristics, few courses obtained even 50% of this optimal score.

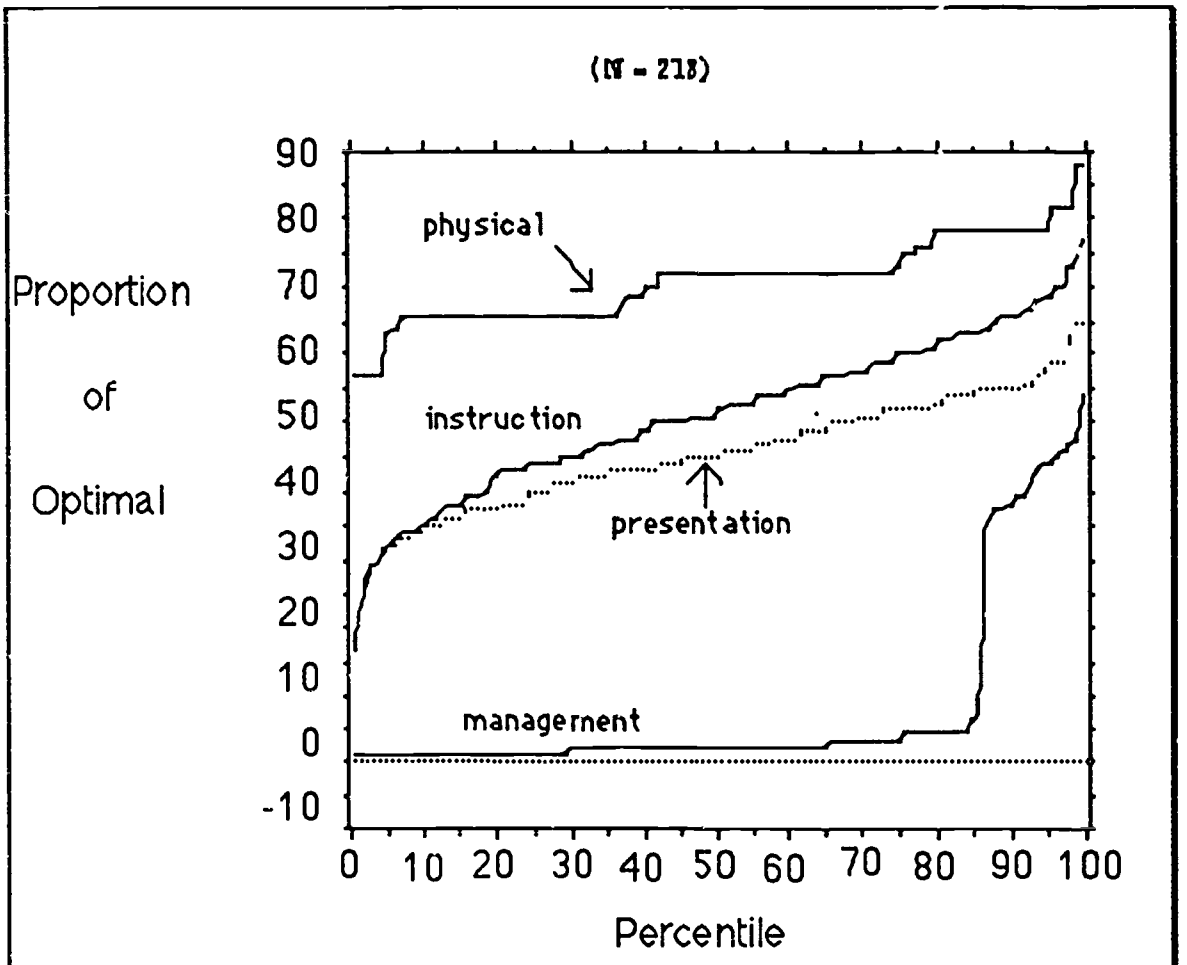

Figure 1: Percentage of Courseware Packages Evaluated that Obtained Specific Levels of Optimality on CCCEM Subscores (Ogive Curves)

Instructional Techniques: A variety of instructional techniques were exhibited by this courseware. Fully 81.6% used page turning during more than 50% of the course duration, while 79.3% made about the same use of tutorials (Figure 2). However, "tried and true" instructional techniques that are usually associated with computer-based learning materials rarely were used an extensive proportion of the time: drill and practice (used extensively 9.4% of the time), concurrent training (7.5%), problem solving (2.4%), modeling (4.3%), and gaming (4.3%). Simulation, as an instructional technique, was used with more frequency, occurring a significant or extensive proportion of the time in 42.4% of the lessons. It is interesting to note that, although instructional computing is much ballyhooed for innovation in education and training, very little use of innovative instruction occurs in this courseware sample. The data indicate that these courses tend to imitate course material

presented in other, more traditional ways, such as lectures or textbooks. Apparently the ballyhoo concerns simply the use of the technology, rather than how the technology is used.

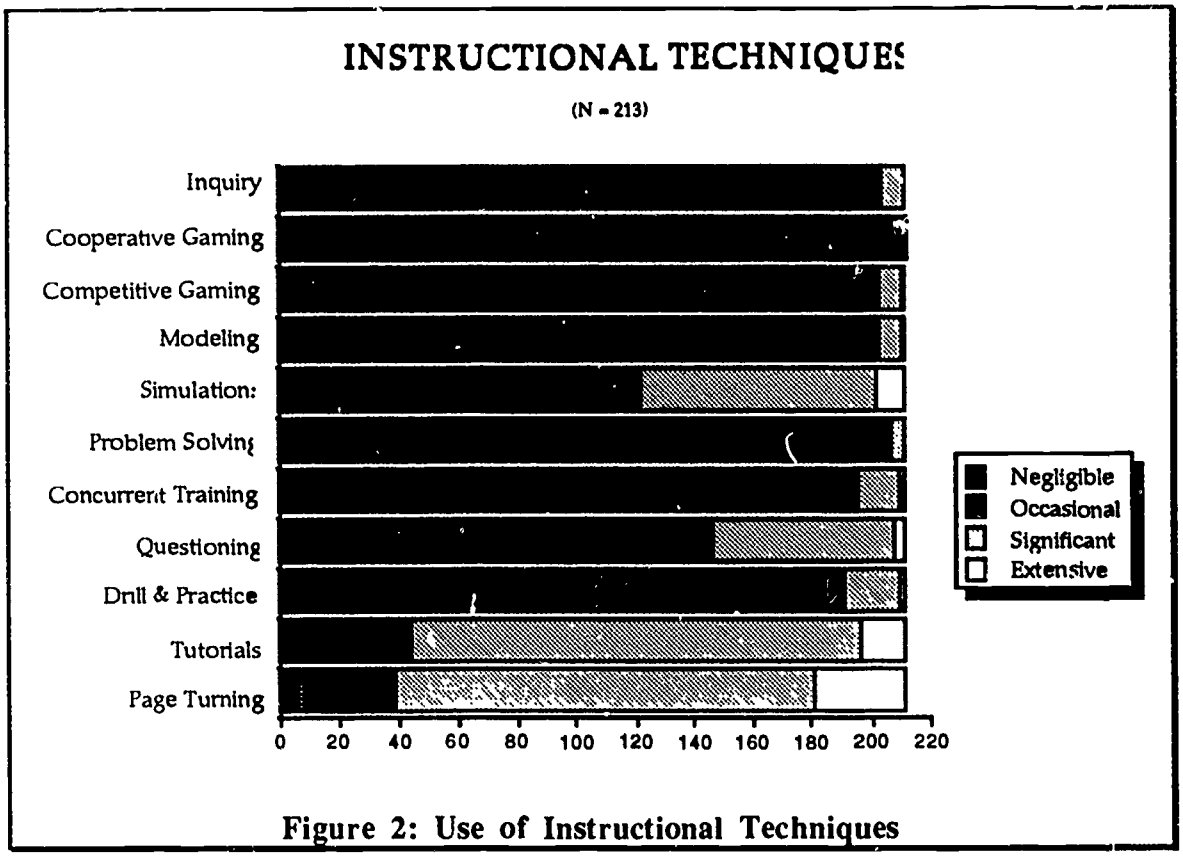

Figure 2: Use of Instructional Techniques

Interaction: This is a crucial component of instructional technology, because it is through the give and take of content material that the learner comprehends the subject matter. Interaction in the CCCEM is measured from two basic approaches as shown in Figure 3:

1. *Computer Output to the learner:* Fully 94.3% of the courses providing feedback to the learner in some way and 47.9% used positive feedback more than half of the course duration. Virtually all courses would prompt the learner in some way when it was necessary to continue the lesson.
2. *Learner input to the computer:* On-line help was available in 39% of the cases, with only 15.4% providing context-sensitive help. This indicates a relatively insignificant use of the technology for the "give and take" inherent in a teaching/learning environment and suggests a marginal use of basic instructional design practice.

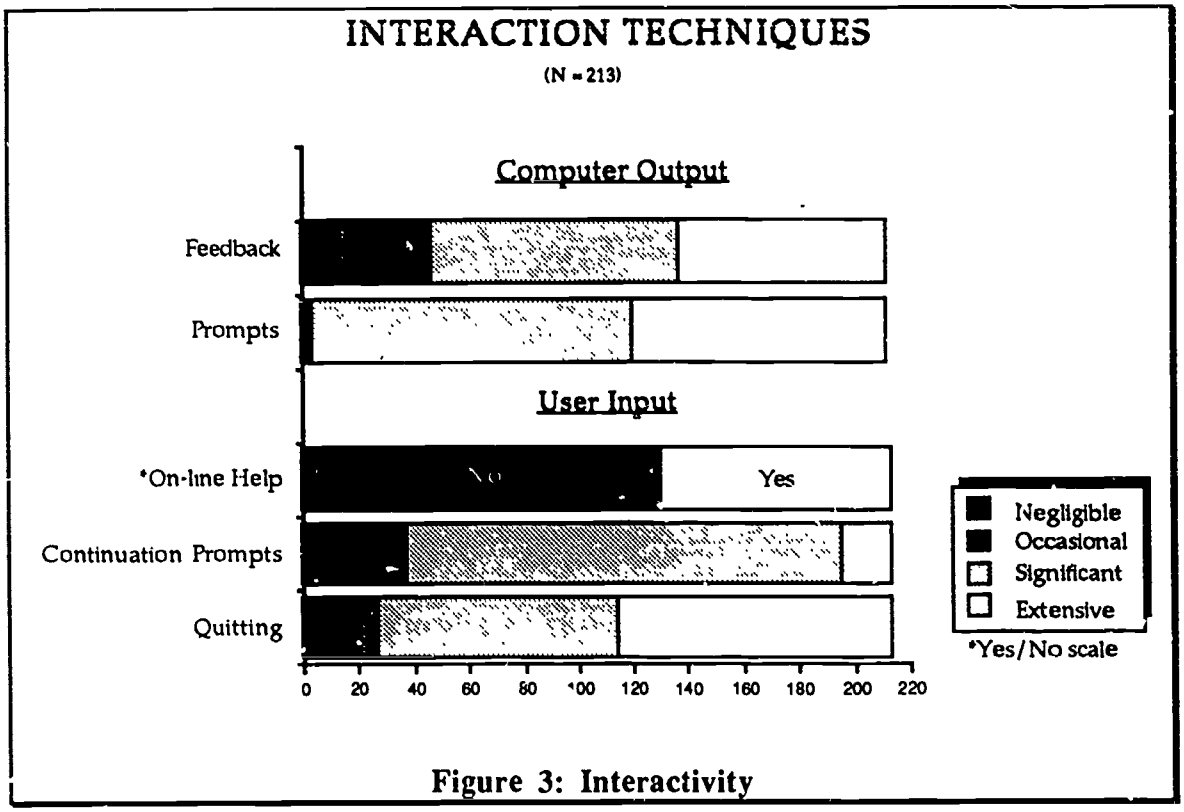

Figure 3: Interactivity

Questioning: Questioning during the lesson can serve to reinforce learning by providing feedback to the learner. Responses by the student can be recorded and the resultant information can be used to branch or modify future learning activities. Figure 4 shows the prevalence of different questioning techniques. Among the courses evaluated, no questions were asked in 37.7% of the packages. This indicates a very low level of interactivity and displays a fundamental design weakness. The lessons involving questioning limited their formats to traditional testing formats [multiple choice questions (70.7%), matching (23%), and true/false (59.4%)] rather than utilizing the technology's capacity to develop interactive techniques that could allow for higher levels of cognitive processing. Questions requiring higher levels of processing [free-form text responses (26.3%); numeric calculation (39.8%)] were far less common. Additionally, no lessons asked for a graphic response. Reducing these numbers even more; most of the courses asking free-form text questions were produced by a single publisher.

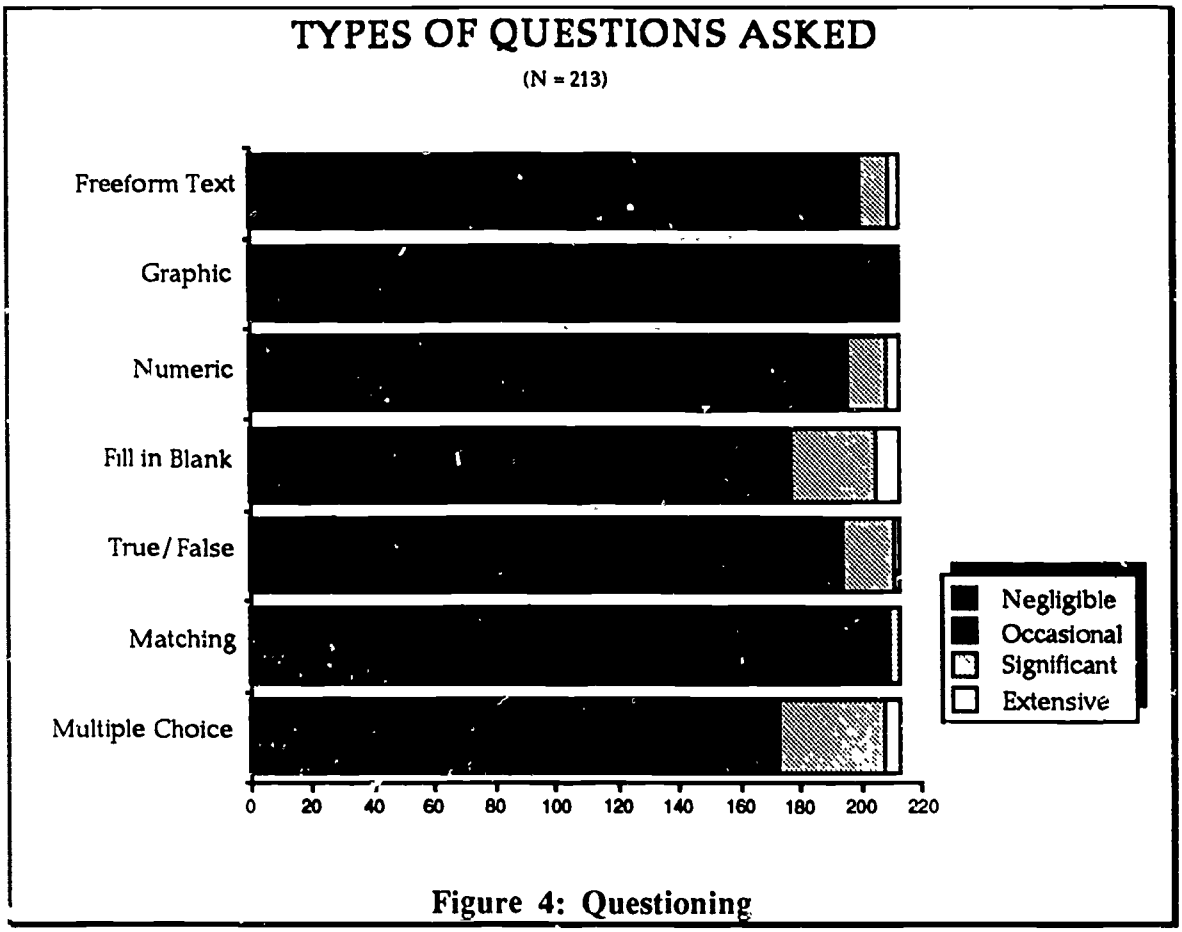

Figure 4: Questioning

Learner control of the lesson: The learner's ability to control a lesson's sequence and timing allows her to feel in control of her own learning process and avoids the feeling of being "force-fed". In a few cases (21.1%) the learner occasionally was forced to wait a period of time before the lesson allowed continuation (Figure 5). However, in only 9.4% of the courses did this occur a significant portion of the time. Forced movement (pacing at a predetermined speed) was even less frequent being present in only 24 courses (11%). About one quarter of the courses which used the forced wait technique allowed the user to override the wait in some manner, as did 33% of the courses which used forced movement.

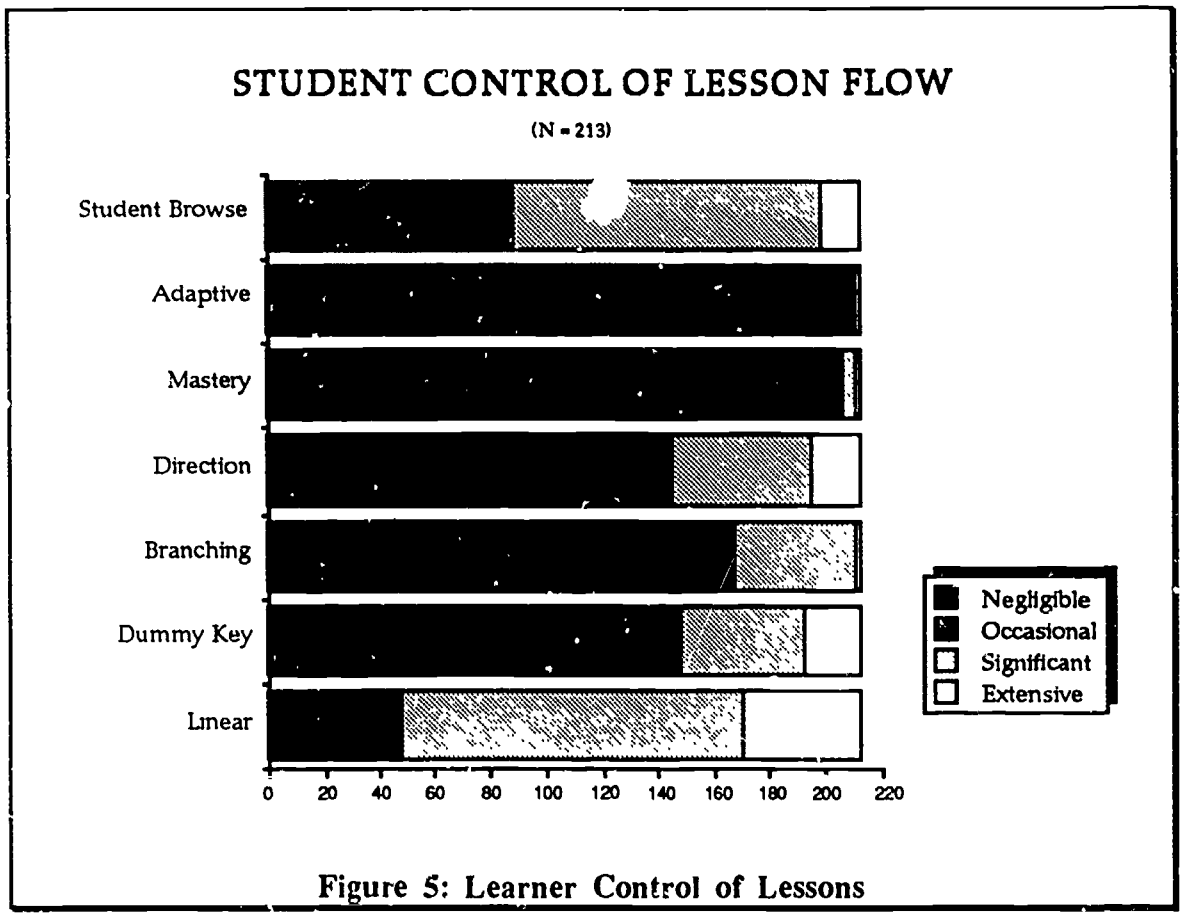

Figure 5: Learner Control of Lessons

Video-based Courseware: CVCEM

Video-based courseware involves an older and more sophisticated technology than CAL. The production and presentation aspects of this courseware were far advanced over those courses evaluated by the CCCEM. Commonly used techniques included: (a) video enhancements such as dramatizations, multiple camera views, varied lighting and dissolve techniques, and (b) instructional techniques such as the regular use of outlines or reinforcers and supplemental materials (e.g. books, seminars). Most of the courses used a single instructor (60%) however, lessons were frequently enhanced by documentaries, dramatizations and demonstrations (Figure 8). Some courses used cueing and 40 used color to provide meaning, The sophisticated use of video effects was prevalent across many courses. Additionally, it was very rare for major presentation problems (sound, production, visual effects, etc.) to occur for any of this courseware. For the video-based courseware it is extremely difficult to separate instruction from production since the latter is frequently used to enhance the former.

Planning and Objectives It is clear from Figure 6 that planning, testing and course management are almost totally lacking for this courseware. Only seven (2.8%) of the courses had any management systems, and task analyses were conducted for fewer than 20% of the courses evaluated. Most of the planning effort was limited to entry level skills and cognitive outcomes as Figure 6 depicts. This clearly shows a fundamental design flaw that appears to prevail among this type of courseware.

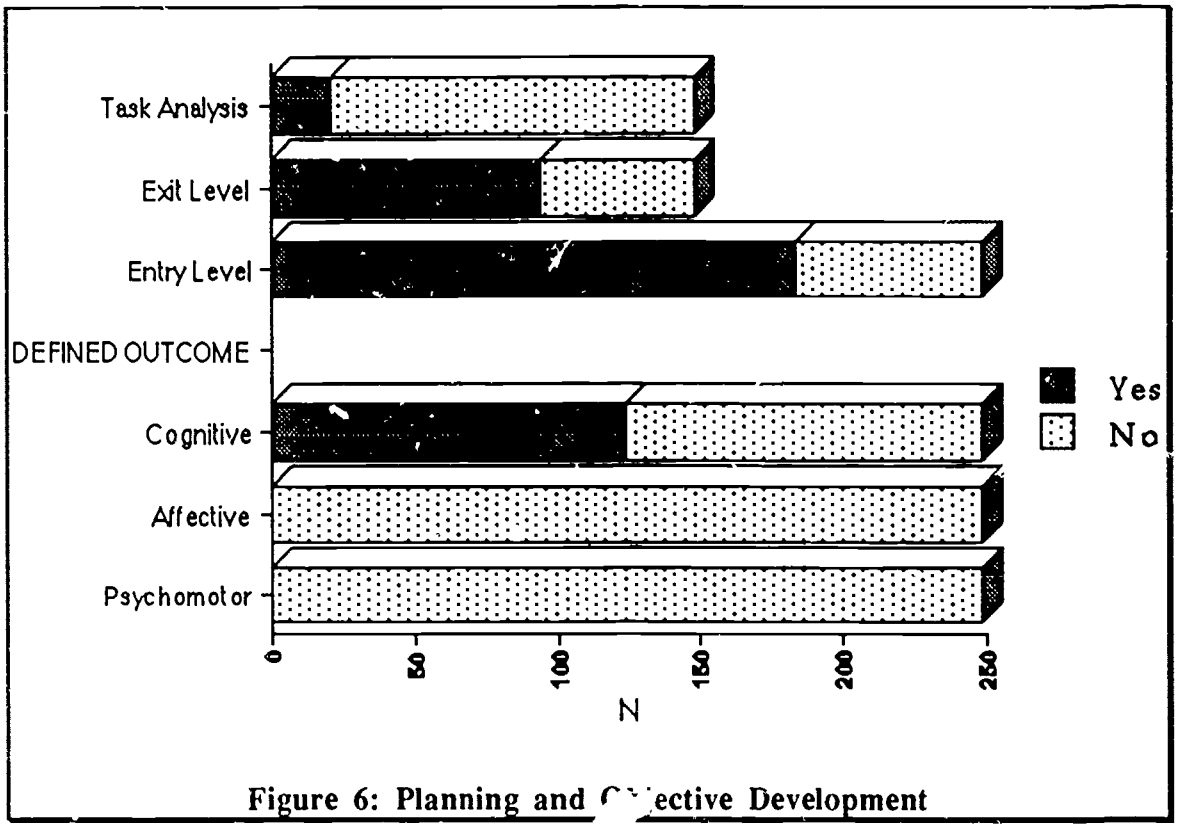

Figure 6: Planning and Objective Development

Measurement & Testing It is interesting how few of the video-based courses measured any student outcomes. As Figure 7 shows, only 12% of the courses that defined outcomes bothered to measure their attainment. This again suggests a basic lack of instructional planning on the part of courseware developers and represents a fundamental defect in courseware design.

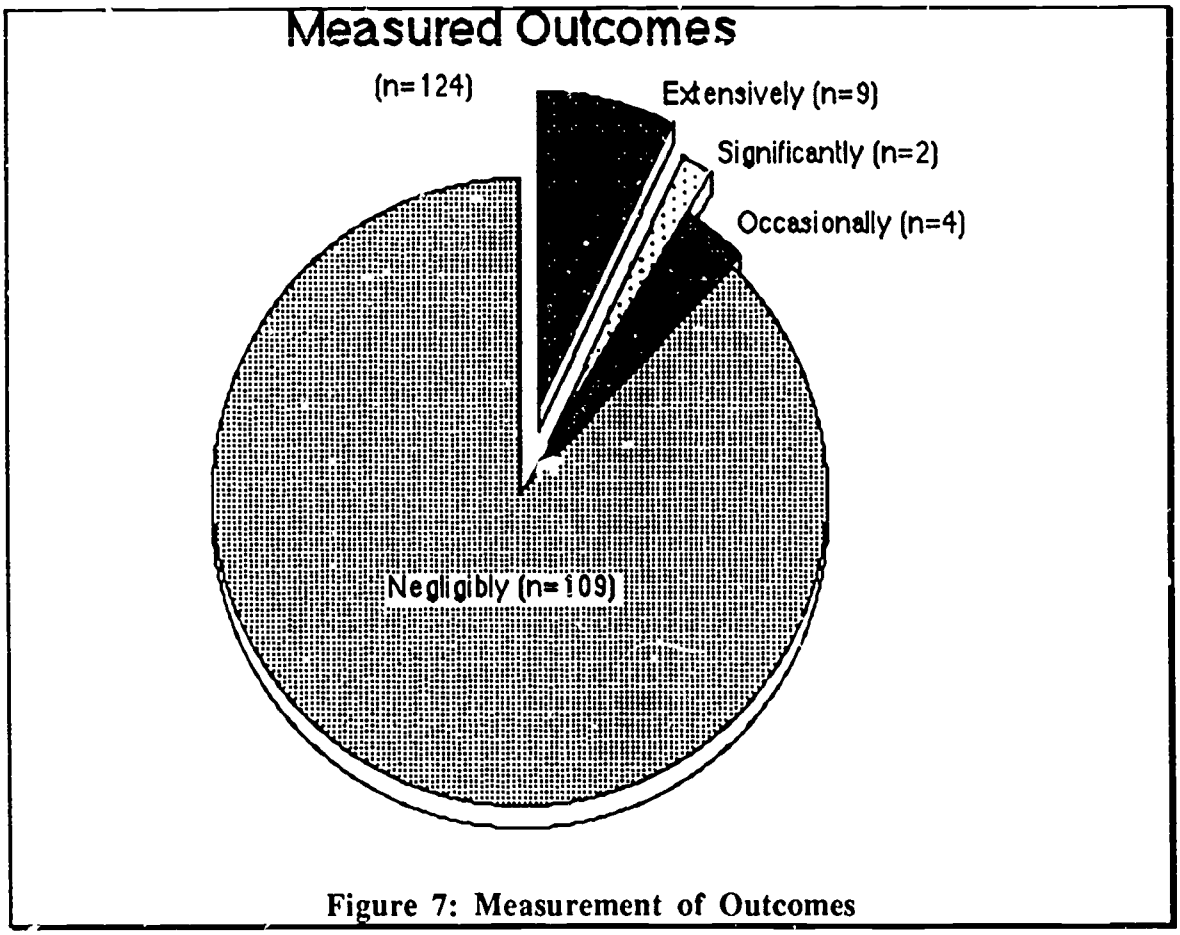

Figure 7: Measurement of Outcomes

Presentation and Production Figure 8 shows that several different types of presentation techniques were used extensively by video-based courseware. In particular, Documentaries, Dramatizations and Demonstrations were commonly used as was the traditional lecture approach. Figure 8 also makes it clear that little use of sophisticated computer technology was found. For example, fewer than 5% of these courses used simulations with any frequency. Again, the capacity of this interactive technology was clearly underutilized by courseware developers.

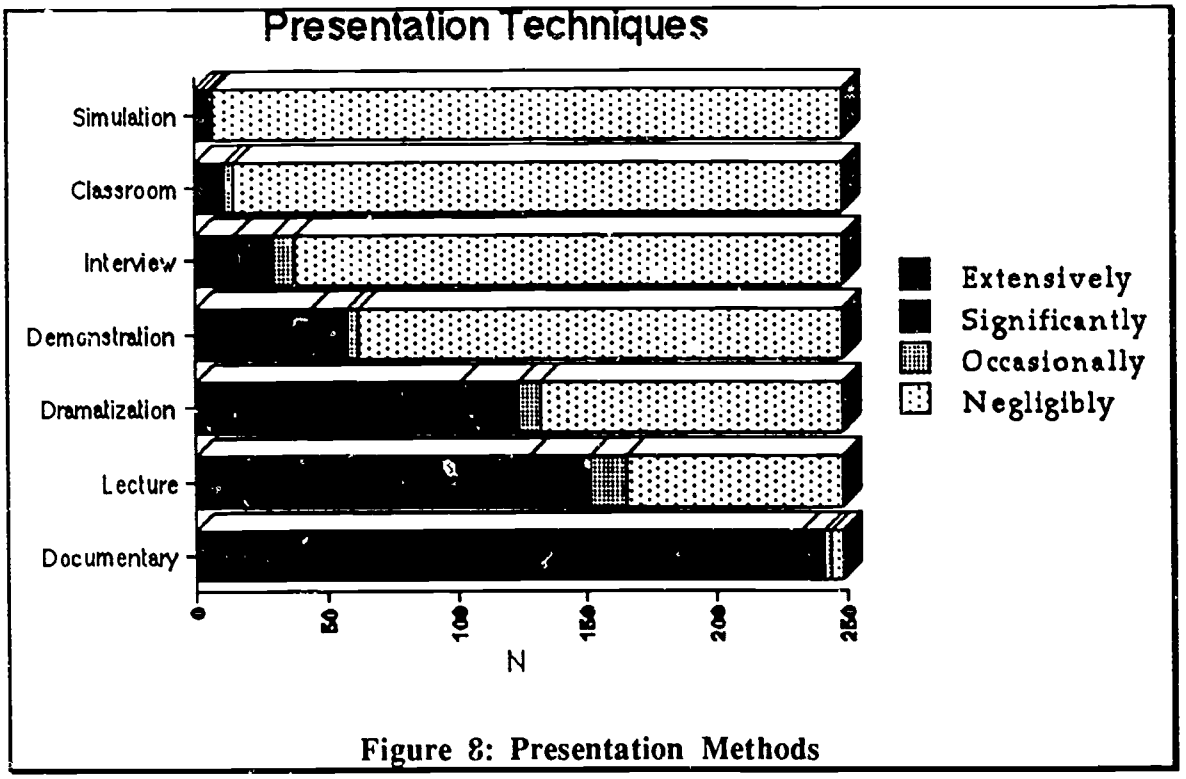

Figure 8: Presentation Methods

Subjective Evaluations Following the completion of a course's evaluation, each evaluator was required to make an overall estimate of each course's **accuracy, depth** and **breadth** of content coverage. Figure 9 shows the positive bias that associated with these variables for the CVCEM. The vast majority of evaluations produced scores of 7 or above on a 10 point scale. Additionally, reliability was sadly lacking for these synthesized evaluations. Although scores for content assessment attained reasonable interrater agreement estimates (.81) in Study 2, their ICC reliability estimates ($r = .27$) shows their inability to discriminate among different courseware [Micceri, 1989]. This small subset of subjective overview contained within the CVCEM appear to be as highly unreliable as typical software evaluations when compared to the objective scores obtained through the use of the CVCEM descriptive and qualitative scores. This further supports the futility of attempting to reliably produce overall ratings for such complex phenomena as computer-assisted or video-based courseware.

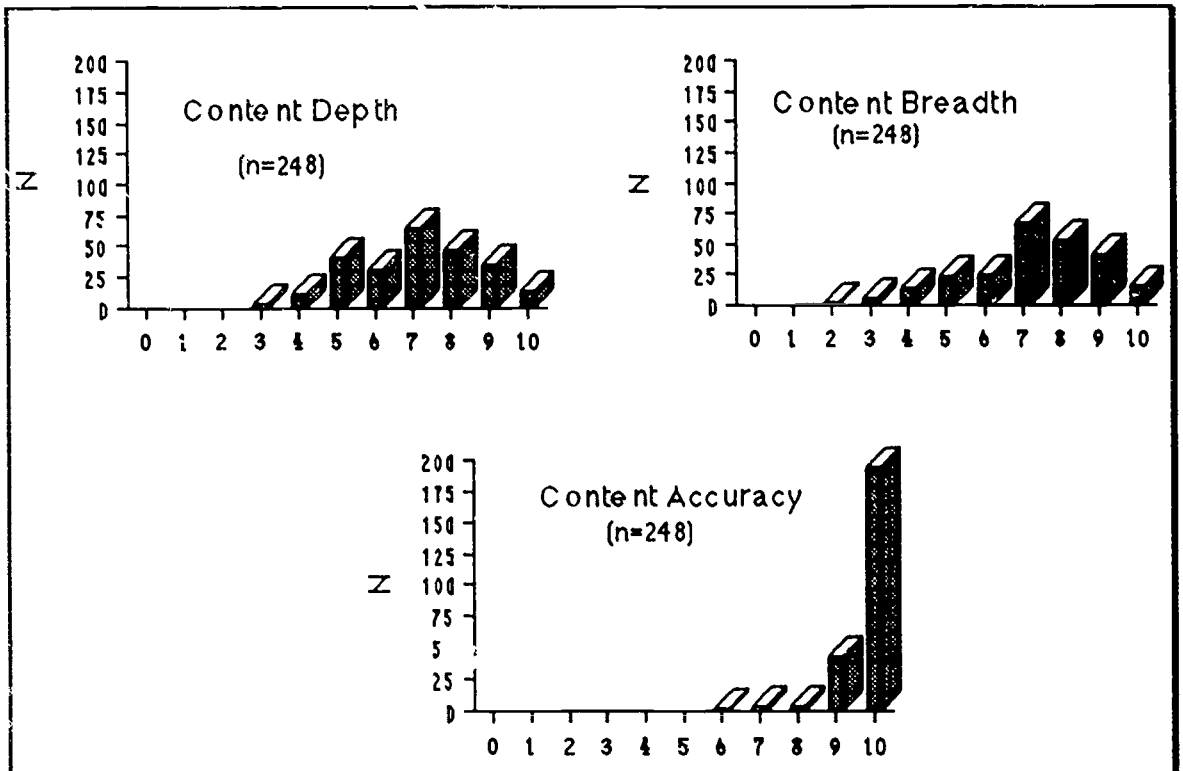

Figure 9: Evaluator's Content Ratings on a 10 Point Scale

CONCLUSIONS AND RECOMMENDATIONS

Today's purchaser of computer or video courseware is often bombarded by seductive visual and auditory stimuli designed to reduce their ability for rational decision making. Unfortunately, it appears that this phenomenon also invalidates most software evaluations since: "... many evaluations also suffer from what might be termed 'The Seductive Nature of Technology'. That is, evaluators, who frequently spend long hours in front of a computer screen viewing familiar subject matter may pay more attention to the presentation than to the instructional aspects of the courseware." (Micceri, Pritchard & Barrett, 1989) As a result of this, many reviews and evaluations of technology-mediated instruction are probably more comparable to theatrical reviews and critiques that to the scientific measures we have come to expect in education. Anyone who has seen Siskell and Ebert do battle is familiar with the various ways two different people can perceive the same product. One should expect that the use of an underlying structure of pedagogically effective methodologies would reduce this effect in reviews of courseware. The two G-studies reported here suggest that the CCCEM and CVCEM models avoid much of this problem and appear sufficiently reliable to provide

consumers with objective evaluations that may be used to compare similar courseware. This is a major step forward in the evaluation of educational technology.

Descriptive information on the courses evaluated provides some interesting findings. For instance, few CAL courses attained even 50 percent of the optimal score defined by the qualitative scaling, itself a considerable reduction from the possible score. The limited scores attained by this courseware suggest that developers fail to take advantage of the many capabilities inherent in these rich technologically enhanced learning environments. Three important areas show particular shortages: (1) management system capacities - where record keeping is generally minimal; (2) instructor's control of lessons - where instructors have almost no control; and (3) the student's flexibility in moving through lessons, which is usually a poor mimicry of microfiche. This last item may be ameliorated somewhat in the future, at least for CAL packages, by the current high level of interest in "hypermedia" as an instructional form.

The video-based courseware also exhibits many of the same problems common to most CAL courseware. Extensive use of sophisticated Presentation and Production techniques and very limited use of the almost limitless pedagogical capacities internal to the technology appear to characterize most examples of this courseware. Additionally, management and planning were almost totally lacking in the video-based courses evaluated.

Publishers of technology-mediated courseware need to recognize that the simple transfer of traditional instruction to the dynamic instructional media described here, perhaps with the addition of a few "bells and whistles" to maintain student interest, does not necessarily imply improved learning. Courseware should be designed from the very beginning with the full potential of the technology in mind. The only way to assure improvement in this is for consumers to only purchase only high quality well designed courseware. In order for this to occur, objective and relevant evaluations that may be compared across courseware appear a necessity. Thus, efforts such as that of CITAR, to develop objective evaluation models, are important developmental steps in this rapidly evolving field.

References

Micceri, T. (1988). Estimating the Reliability of the CITAR Computer Courseware Evaluation System. *ERIC* (Document No. ED 295 971, TM 011772).

Micceri, T. (1989). Estimating the Reliability of the CITAR Video-Based Courseware Evaluation Model. Internal technical report: CITAR, Tampa: University of South Florida College of Engineering.

Micceri, T., Pritchard, W.H., Jr., & Barrett, A.J. (1989). Must Computer Courseware Evaluation be Totally Subjective? The Development of an Objective CAL Evaluation Model., The British Journal of Educational Technology, 20:2, p. 120-128.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and the generalizability of data collected in observation studies. Psychological Bulletin, 86, 376-390.

Peterson, D., Micceri, T. & Smith, B.O. (1985). Measurement of Teacher Performance: A Study in Instrument Development. , Teacher and Teacher Education, 1:1.

Pritchard, W.H., Jr., Micceri, T, & Barrett, A. J. (in press). A Review of Computer-Based Training Materials: The Current State of the Art (Instruction and Interaction), Educational Technology.

Shrout, P. E. & Fleiss, J. L. (1979). Intraclass correlation: Uses in assessing rater reliability. Psychological Bulletin, 81, 420-428

END

U.S. Dept. of Education

Office of Education
Research and
Improvement (OERI)

ERIC

Date Filmed

March 21,1991