The Effect of Electronic Evaluation Tools on Types and Magnitudes of Learning

Suzanne Daningburg
Richard F. Schmid

Abstract: The effects of use and non-use of a particular electronic formative evaluation tool, the Program Evaluation Analysis Computer (PEAC) system on the cognitive, affective and evaluative aspects of college students' perception of an educational television (ETV) program emphasizing cognitive learning were investigated. A specific real-time evaluation question was used for Study 1 and a more global question for Study 2. Scores for theshort answer test assessing content recall in Study 2 increased from pretest to posttest for non-users and decreased for users, suggesting that PEAC usage hindered recall involving comprehension. No other PEAC effects were found. The results suggest that while electronic evaluation techniques may be well suited for evaluations involving affective aspects of ETV programs, they may not provide useful evaluation information when objectives emphasize cognitive change, and may actually interfere with some types of learning. Formative evaluation factors Influencing categories of learning are discussed.

Resume: lei, nous avons etudle les effets qu'ont 1'utllsation et la non-utilisation d'un outil d'evaluatlon formative specifique, le systeme devaluation de programme «PEAC», sur les aspects cognitifs, affectifs et evaluatifs de la perception qu'ont les etudiantsde niveau collegial du programmede television educative«ETV» qul porte sur l'apprentlssage cognitlf. Une question specifique en temps reel a et6 utllisee dans l'etude No.l et une question plus globale dans l'etude No. 2. Dans l'etude No.2, les scores obtenus, en reponse aux questions a reponses breves portant sur 1'evaluatlon du contenu retenu, ont augmentedu pretest au post-test pour les non-utilisateurs mals dlminue pour les utilisateurs. Ce resultat laisse supposer que 1'utllisatlon du PEAC est un obstacle a la retention qui demande une comprehension. Aucun autre effet reli6 au PEAC n'a ete releve. Les resultats suggerent que, blen que les techniques d'evaluatlon informatisees peuvent etre parfaitement utiles dans revaluation des program mes ETV com portant des aspects affectifs, elles ne peuvent peut-etre pas donner une evaluation valable quand les objectifs mettent en valeur lechangement cognitif, Elles pourraient memeentraver certaines formes d'apprentissage. Les facteurs devaluation formative pouvant influencer les genres d'apprentissage y sont egalement trait6s.

In the field of educational technology, the breadth and potential sophistication of media applications has dramatically increased the need for formative evaluation of educational and training processes and products. A number of works have delved into the history (Cambre, 1981), methods (Weston, 1986),

Canadian Journal of Educational Communication, VOL. 23, NO, 1, PAGES 3- 18. ISSN 0710-4340

definitions (Scriven, 1967; Dick, 1980; Gooler, 1980) and tools (Millard, 1992) of formative evaluation. While this literature provides a solid descriptive and methodological foundation for classic implementation of formative evaluation, dramatic changes in instructional design have necessitated the conscious re-examination of the evaluation process. For example, front end analytic approaches such as performance and needs analyses have expanded the purview of educational technology to include all aspects of the target system, not just the student and content (Rosenberg, 1990). Techniques and tools as diverse as rapid prototyping and small format video are bringing designers, developers and end users closer and closer together, necessitating the evaluation of the entire learning system (Daningburg & Schmid, 1988). Alongside these systems, powerful, electronically-based design and evaluation instruments have been developed which are fundamentally changing how we conceptualize and develop learning products and engineer their use.

The purpose of the present studies was to focus on these new electronic evaluation tools, and examine their utility in media development. Measurement tools such as computer-based, on-line audience response systems are now a standard part of educational television production. These electronic systems provide designers and researchers with virtually instant access to viewer reactions to the content being presented, and are demonstrably powerful in yielding accurate, affective information on viewer preferences (Baggaley, 1982). These tools have encountered remarkable success at providing information with face and content validity (i.e., samples from the target population tell us exactly what they think moment by moment, and revisions are made accordingly). However, the value and effect of these tools in terms of empirical or construct validity are more difficult to ascertain, and virtually no empirical research has addressed these issues. Of general concern is the extent to which these tools, by their very presence, alter the phenomenon they are measuring. The validity of their output, as utilized in the instructional design process, thus warrants scrutiny.

Electronic Measurement Methodologies

In order to understand the potential role of electronic measurement tools in instructional design, it is first necessary to describe them. Typically, these systems electronically gather second-by-second reactions from individual viewers to single evaluation questions such as "How interesting is the program?" The data, stored in hand-held units which resemble small pocket calculators, consist of reactions from one to many viewers evaluating simultaneously. These data are dumped into a central microcomputer, aggregated and aligned with each moment of the program. The output, including the superimposition of real-time viewer responses over an entire program or segment, is specially designed for meaningful input to program production decisions (Nickerson, 1979). Fine-grain analyses are possible, with visual representations of viewer involvement and/or enjoyment directly linked to program dynamics (e.g., specific script lines, appearances of a given personality, scene or format change). By capturing an aspect of the viewer's immediate reaction, these systems allow for remarkable precision in quality and

degree of evaluative detail. Their key strength is that viewer reactions are time-locked to the programming which evoked those reactions. Baggaley (1982) notes, "Audience reaction can shift and change from one moment to the next, and overall reactions to a programme can be due to an isolated moment within it which even its producer cannot predict" (p. 70). These systems overcome the obtrusiveness of inserted oral or written questions (which require actual pauses in the program) and the lack of specificity inherent in a reliance on post-presentation questions.

A prime example of this type of electronic evaluation tool is the Program Evaluation Analysis Computer (PEAC) system. We selected the PEAC system as a representative tool for these studies because it has been used in a wide variety of TV-based projects (Corporation for Public Broadcasting [CPB], 1981; Radio-Quebec, 1985, 1984; Baggaley, 1982) and because it is similar to other popular electronic techniques. (See Millard, 1992, for a complete description of such tools.) While the PEAC system has been used more for formative evaluation of commercial productions than for educational materials, the growing trend to more sophisticated tools makes the use of systems such as the PEAC more and more desirable. By examining the effects of a prototypic system, some general principles might emerge regarding formative evaluation.

Electronic Evaluation and Learning System Design

Automated evaluative tools such as the PEAC system have proven useful in assessing affective variables such as viewers' interest, positive (or negative) reaction, recognition and/or empathy for actors/products, and so on (e.g., CPB, 1981). However, instructional designers and teachers are also concerned, if not preoccupied, with cognitive objectives (Clark, 1992; Daningburg & Schmid, 1988; Romiszowski, 1981). The use of audience response systems within an educational context raises two related questions. First, does their current use in assessing affective variables interfere with or distract designers from the assessment of equally or more important objectives, such as cognitive ones? Second, can these tools also be used to enhance the effectiveness and/or efficiency of the development process for cognitive content? The issue of whether electronic measurement usage affects viewers' cognitive processing is important because formative evaluators who use continuous opinion measures during viewing tend to ask comprehension questions after viewing. The evaluators usually assume a negligible interference effect (Millard, 1992). We know of no empirical research which has studied the possible obtrusiveness or interference of electronic evaluation techniques with the cognitive processes involved in learning. (See Baggaley, 1987, for a review of various continuous response factors.)

Evaluation Process and Instructional Product

In addition to examining evaluation factors within the educational production system, general variables which have some influence over the final audience response need to be highlighted. Daningburg and Schmid (1988) examined the interaction of three general components of the evaluation process of educational tools: the producer; the teacher and the learner; and the effect of this interaction

on the instructional product. Within their model, five factors were identified which seem to play an important role in learning effects: assessment objectives; individual differences; content familiarity; mathemagenic activities; and attitude and motivation. Each of these variables merits a brief comment on the context of its relation to the evaluation process.

Although the *sine qua non* of assessment objectives is that they should evaluate what they claim to evaluate, electronic techniques may not achieve this essential standard. Empirical demonstrations which establish the validity of tool usage on various types of objectives are essential. The second factor, individual differences, has long been recognized as a key determinant in the success of instruction (Bracht, 1970; Bracht & Glass, 1968; Cronbach & Snow, 1977; Snow, 1978). Notwithstanding the diminished role of Aptitude by Treatment Interaction (ATI) research in educational circles (McCain, Short & Stewin, 1991), individual differences must be taken into account in evaluating a medium increasingly targeting specific, well-defined populations and segments thereof. Third, content familiarity is a well-recognized cornerstone of the instructional design process, requiring sharp focus on learners' prior knowledge and ability. Unfortunately, familiarity is complicated by both interest level and presentation format of content (e.g., boring or fascinating; easy or difficult). The fourth factor consists of mathemagenic activities, or those student behaviors relevant to the achievement of instructional objectives in specified situations or places. Rothkopfs early work (1966; 1970) is still relevant today: he argued that by drawing attention to certain aspects of instructional content, questions asked prior to, during, or after a lesson differentially shape the student's processing and learning. Finally, attitude and motivation have an impact on learners' attention and effort in learning. Programs must have affective goals and criteria to measure them because it is difficult for any but the most highly trained professional evaluator to separate a cognitive response to a given event (or instruction, etc.) from a feeling or an opinion about it. Each of these five variables, central to the learning process, may be positively or negatively affected by electronic evaluation techniques. The purpose of this research was to assess these effects.

The Studies

As noted above, one of the main contributions of electronic evaluation tools to the process of formative evaluation is that they significantly reduce many of the obtrusive aspects of the evaluative process (Nickerson, 1979). The question addressed by these studies is what remaining (or new) obtrusiveness exists, first in the cognitive domain where instructional objectives usually reside, and in the affective domain, where the PEAC system is most often applied. In order to examine this question of obtrusiveness on the learning process, two levels of PEAC usage were utilized: learners using the technique, and learners not using it. Assuming that attention is selective, the key to assessing differences between the groups as a result of treatment depends upon the sensitivity of the dependent measures to effects resulting both from evaluation task demands (critical information to which attention is drawn) and from aspects not directly addressed in

the evaluation process (incidental information) (Anderson, 1972). These studies explored the relationship between the task demands imposed by the PEAC system and the specific content being evaluated. For example, if the evaluator were to ask learners to assess the visual quality of the program, the increased attentiveness of viewers to that aspect of the content would likely enhance their subsequent recall of it, probably to the detriment of other aspects (Rothkopf, 1970; Watts & Anderson, 1971). While this attention factor is fairly obvious, the process represents a dilemma for the evaluator. The more precise, and thus prescriptive and useful the information, the more obtrusive it becomes to other factors, regardless of the measurement technique (PEAC or otherwise). The evaluator is therefore left with measuring either global responses which preserve external validity, or specific questions which by their very nature prime attention to certain aspects while interfering with others. Thus, formative evaluation must both carefully specify precisely what it is that needs to be measured, and recognize that that decision will have an impact on the generalizability of the results.

The studies reported here were exploratory in that they attempted to integrate several areas of inquiry which seemed to be logically related. The major effect expected was that use of the PEAC system would influence cognitive recall in certain conditions. (In the interest of clarity we report the studies in terms of the actual variables used in analyses, rather than in terms of the initial design. Readers wishing complete information are invited to contact the first author for details.)

METHOD

Design

A Pretest-Posttest Control Group fixed design was used, with the Usage (use or non-use of PEAC) factor between subjects. The pretest used to determine content familiarity was used again as the posttest, leaving open the possibility of including a repeated measures comparison. While the same design was used for two populations, described below as Study 1 and Study 2, each population received a different overall evaluation question.

Subjects

The subjects in Study 1 were 55 students from a private urban college. For Study 2,69 students from a public urban college were used. Since subjects' prior knowledge of content was assessed via a pretest rather than manipulated, random assignment of groups was preserved, thus avoiding the limitations associated with an ex post facto design.

Materials

The instructional stimulus consisted of a 27-minute Educational Television (ETV) program entitled "Out of the Mouths of Babes", produced in 1975 by the Canadian Broadcasting Corporation (CBC) and previously aired on the weekly

'The Nature of Things" television series. The program examined the theory that a child's ability to talk is partly innate; that babies are born with the capacity to extract not just words, but rules of grammar, from what they hear spoken around them. This particular program was selected for three reasons. First, its content and level of production were appropriate for college-level students. Second, it was believed that the content would be familiar to some students and unfamiliar to others, allowing an adequate number of learners in each level of content familiarity. Third, the program was considered a good representation of a typical ETV program in that it was of relatively high production quality and it dealt with an academic, though popular subject matter.

The pre-test measures consisted of a two-part prior knowledge cognitive test and an attitude questionnaire. Prior knowledge was assessed by two composite measures, one based on 6 short answer items and the other on 10 multiple choice items. The items tested levels of knowledge, comprehension and application of program content. The attitude measure consisted of 14 scaled Likert-type questions which were related to subjects' attitudes toward the importance and relevance of language acquisition in children. Each attitude item dealt with a discrete topic, and was therefore treated individually.

For the first session, the experimenter orally presented instructions for each of the pre-treatment measures, information about program viewing and a request to complete a post-viewing questionnaire to all groups. For the experimental sessions one week later, the experimenter requested that students watch the program, and gave instructions regarding each of the post-viewing questionnaires. In addition, those groups using the PEAC system were instructed verbally in the use of the PEAC devices.

For Study 1, all subjects were told that the overall question to be used in evaluating the program was: "How effective do you think this program is in demonstrating language acquisition in children?" The question was designed to prompt learners to attend to the academic objective of the presentation, namely, the presentation of instruction about childrens' language acquisition. Subjects using the PEAC system were asked to rate the program in terms of this overall question, using their hand-held devices. Possible responses were: 'Very effective"; "fairly effective"; "not very effective"; and 'Very ineffective". Subjects not using the PEAC system were simply asked to keep the question in mind as they viewed the program.

It was initially intended to use the same overall (real-time) question for both samples. However, during Study 1 it was found that the interaction between this question and the characteristics of the program failed to provide for discriminating evaluation. Nearly all subjects using the PEAC system rated the film as 'Very effective" or "fairly effective" at all times and in fact used their hand-held PEAC devices very infrequently. The decision was made to change the overall question for Study 2. All subjects for Study 2 were told that the overall evaluation question was: "How good do you think th is program is, based on your overall reaction?" This question was similar to the general content evaluation typically used in much of Baggaley's work. Subjects using the PEAC system were asked to: "Kate the

program as very good, fairly good, not very good or poor. Use your own judgment to make the evaluation, based on anything that strikes your overall reaction at any given moment." Once again, subjects not using the PEAC system were asked to keep the question in mind during viewing. It was hoped that the change of question would accomplish two goals. First, the more global nature of this question would encourage students to be more discriminating about different aspects of the program. Second, the simpler, more understandable phrasing would make the task easier and thus encourage more frequent responses. For both studies, in classes using the PEAC system, the questions and possible ratings were written on the blackboard as a reminder to subjects throughout the program.

The post-presentation dependent measures consisted of five parts. The cognitive and affective measures given in the pre-presentation session were re-administered. An additional section was attached to the cognitive component of the posttest. It was thought that this measure might provide a more precise indication of the relative obtrusiveness of the PEAC system. Subjects' own opinion about the effectiveness of the program was sought through 20 program evaluation rating questions. These were different from the attitude measures in that the program evaluation questions tapped opinions about the presentation of this specific program whereas attitude items probed opinions related to the general topic — that is, language acquisition—apart from the ETV program. The items in the program evaluation rating section, designed to measure opinion about the actual program, varied in nature, with some having a short answer format, some a multiple choice format, and others a Likert-type scale format. Finally, 16 demographic questions ascertained information such as age and sex.

Procedure

Study 1 involved three intact classes, two of which made up the experimental group (re = 16 + *n* = 13), thus (n=29), and one control (n=26). Study 2 also involved three intact classes, two of which made up the experimental group (*n* = 23 + *n* = 17), thus (*n* = 40), one control (*n* = 29). Any potentially different learner characteristics were assumed to have been distributed in random fashion, since the three intact classes in each of the studies consisted of different sections of the same course, and were taught by the same instructor. The students had been placed in a given section on the basis of their college's scheduling system, so that confounding group differences were highly unlikely.

At the first session the regular instructor introduced the experimenter, who gave a brief introduction about the study, and distributed large envelopes containing the pre-treatment questionnaires. Students were asked to complete the questionnaires, according to ensuing verbal directives and written instructions. Time limits were imposed for each questionnaire. When all questionnaires had been completed, students returned them to the experimenter. Students were told that the second session would occur in one week.

One week later, all groups were given an identical introduction to the ETV program. Subjects were informed that they would be asked to answer questions after viewing. In addition, experimental groups (i.e., those using the PEAC system) were given oral instructions on the use of the PEAC system (including an explanation of the overall real-time question as mentioned above). The program was shown, after which three questionnaires were distributed to each student. Once again, students were asked to complete the questionnaires according to verbal directives and written instructions, and a time limit was imposed. When finished, students returned the questionnaires and the experimenter answered any questions about the study.

RESULTS

Scoring Procedures for Dependent Variables

The two cognitive tests, the short answer format and the multiple choice format, were used as measures of content familiarity in the pretest. The six short answer items were assigned 2 points for each correct response, and only one point was given for partially correct answers. An inter-rater reliability coefficient of .91 was obtained for the short answer items. The ten multiple choice items were assigned one point for each correct response. Each item in the attitude questionnaire was coded individually, with a value from one to five to each of the responses.

The cognitive posttest consisted of the same two measures that were used in the pre-treatment session, and were scored the same way. An additional test, designed to further explore the relation between use of the PEAC system and different types of learning, consisted of 11 fill-in-the-blank type questions related to incidental learning. A total of 26 points was possible, with partial credit given to responses containing fewer than the total concepts or words required. The attitude posttest was identical to that of the pre-treatment sessions and was scored in the same manner. Demographic information was also collected and tabulated.

Tests of Assumption

Due to the need to employ intact classes within each study, equivalence of groups was assessed prior to the analyses of effects. Analyses of variance (ANOVAs), completed on the three classes of each study for all of the pre-session dependent measuring instruments yielded no differences, thus verifying class equivalence within each study.

Study 1

Study 1 involved data from the private college sample only. The sample from the public college (Study 2) was significantly different from its private counterpart on pretest performances, thus amplifying the need for separation of the two groups.

Means and standard deviations for the three cognitive variables—short answer, multiple choice and incidental learning scores—are listed in Table 1. Results of multivariate and univariate analyses of variance yielded no significant results.

TABLE 1
Cognitive and Aptitude Measures for Study 1

Criteria	With PEAC			Without PEAC		
	X	SD	n	X	SD	n
Pretest						
Short answer	3.83	1.42	29	3.50	1.45	26
Multiple choice	3.10	1.57	29	2.81	1.50	26
Posttest						
Short answer	4.38	2.24	29	3.77	1.56	26
Multiple choice	7.90	1.52	29	8.00	1.50	26
Incidental	15.59	2.77	29	16.12	3.30	26

For the attitude items, each of which dealt with different components of the program and content, Chi-square and Mann-Whitney U tests were used. No significant differences between use and non-use groups were found for either the attitudes expressed on the pretest or those expressed on the posttest. Despite the ordinal nature of scores on the attitude items, a repeated measures analysis of variance was conducted in order to determine whether or not any changes occurred from the pretest to the posttest within each group. A total of 7 cases were deleted from the experimental group, while 2 cases were deleted from the control group due to missing data. None of the response distributions was statistically significantly skewed. Results from this analysis yielded results similar to the non-parametric analyses mentioned above. For example, both use and non-use groups expressed the opinion that television was a desirable medium by which to learn about topics like language acquisition; and both groups also shared the attitude that the documentary format was not necessarily appealing, after viewing the ETV program.

No significant differences were found for any of the program evaluation rating questions (which probed viewers' opinion about the particular program itself rather than their attitude about the topic and television in general), using Chi-Square tests.

Study 2

The results of Study 2 were based on data from students at the public college. The specific tests dealing with cognitive recall, attitudinal responses, and

program evaluation ratings were identical in all respects to those of Study 1.

Means and standard deviations for the three cognitive variables — short answer, multiple choice and incidental recall — are listed in Table 2.

TABLE 2
Cognitive and Aptitude Measures for Study 2

Criteria	With PEAC			Without PEAC		
	X	SD	n	X	SD	n
Pretest						
Short answer	2.65	1.48	40	2.35	1.52	29
Multiple choice	1.90	1.15	40	2.36	1.25	29
Posttest						
Short answer	2.05	1.65	40	3.41	1.99	29
Multiple choice	5.60	1.78	40	6.18	1.49	29
Incidental	13.23	4.46	40	13.66	5.02	29

A multivariate analysis of variance (MANOVA), using the short answer and the multiple choice scores of the pretest and the posttest as dependent variables, resulted in a significant Retelling test ($F(l,63)=3.73, p=.009$) for the main effect of PEAC usage. No other overall effects were significant. Within the PEAC effect, univariate analyses showed the short answer scores to be significant ($F(l,66)=10.26, p=.002$\ while the multiple choice scores only approached significance ($F(l,66)=3.5$, $p=.066$), both on the posttest only. Table 3 represents the multivariate and univariate comparisons for these dependent measures.

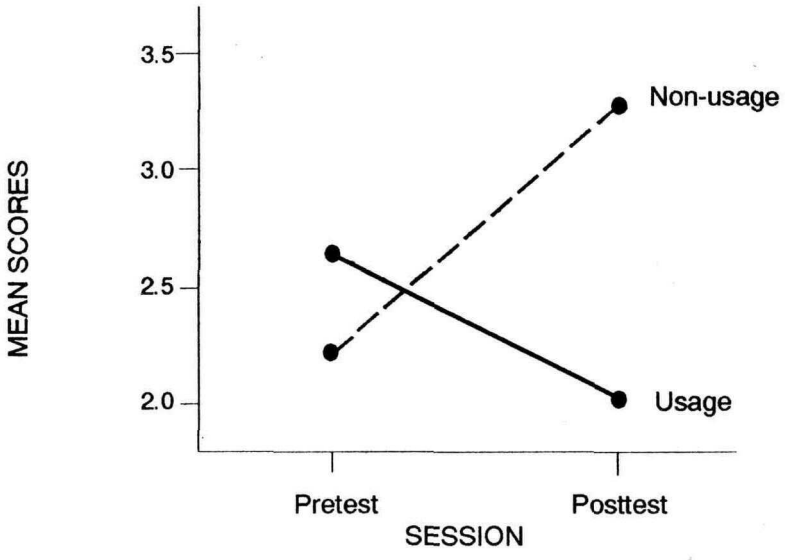
These analyses showed that, based on the short answer posttest scores, non-users of PEAC outperformed users of PEAC.

The above MANOVAs provided essential information regarding the central hypothesis of the study, namely, that use of the PEAC system influences cognitive recall. In order to further examine the nature of this influence, overall pretest performance was used as a baseline and compared to posttest performance. Thus, a repeated measures analysis of variance with the short answer pretest and posttest scores as the repeated measures and PEAC usage as the treatment factor was performed. The results produced a significant interaction ($F(1,66)=11.64$, $p=.001$). While both experimental and control groups performed at the same level on the pretest, the control group, which was not distracted by the PEAC evaluation, improved significantly (Tukey, $q(67)=4.06$,/><.01) and the experimental group actually performed slightly worse on the posttest. On the posttest comparison, the control group also performed significantly better than the experimental group (Tukey, $q(67)=5.56, p<.01$). A graphic representation of this interaction appears in Figure 1.

TABLE 3
Multivariate and Univariate Tests on Short Answer Pretest (SAPRE), Multiple Choice Pretest (MCPRE), Short Answer Posttest (SAPOST), and Multiple Choice Posttest (MCPOST) Scores for Study

Effect	Multivariate test (Hotelling's)			Univariate tests Error Mean Square: SAPRE 2.27, MCPRE 1.39, SAPOST 3.24, MCPOST 2.84			
	F	df	*p*		F	df	P
PEAC	3.73	(1,63)	.009	SAPRE	.63	(1,66)	.429
				MCPRE	1.89	(1,66)	.174
				SAPOST	10.26	(1,66)	.002
				MCPOST	3.50	(1,66)	.066

Figure 1.
Mean Scores on Short Answer Pretest and Posttest for Usage and Non-Usage of PEAC for Study 2.



As in Study 1, attitude items were examined individually using non-parametric analyses. Mann-Whitney U tests on each item showed that on the pretest the PEAC usage group attached significantly more value to the importance of language acquisition in children than did the non-use group. No significant differences between groups were found for attitudes indicated on the posttest.

In order to assess changes in attitudes from pretest to posttest within groups, a repeated measures analysis of variance was performed. Due to missing data, 12 cases were deleted from the experimental group, while 6 cases were deleted from the control group. None of the response distributions was statistically significantly skewed. As in Study 1, both the use and the non-use groups seemed to share certain attitudes. For example, two items indicated that after viewing the program, students expressed a more negative attitude toward the importance and/or relevance of language acquisition in children.

Results from the Chi-Square tests performed on the program evaluation rating questions showed no overall differences between groups.

DISCUSSION

The study's central hypothesis, which stated that the use of an electronic evaluation technology would influence cognitive learning, was supported for content recall when a general real-time evaluation question was used. What also emerged, however, was the realization that the validity of the evaluation process and outcomes is largely determined by the context within which the evaluation is carried out. A given evaluation tool may be appropriate for certain situations, and inappropriate for others. What is therefore needed are guidelines for determining when and where electronic tools such as thePEAC system are useful.

This discussion will return to the five principle contextual factors cited from Daningburg and Schmid (1988) above — assessment objectives, individual differences, content familiarity, mathemagenic activities, and attitude and motivation — to interpret the results.

Assessment Objectives

In looking at assessment objectives, two critical questions must be addressed. First, what are the intended learning outcomes of aproduction? And second, what is it that is being evaluated? The results of these studies suggest the direct involvement of these contextual factors with the task demands placed on learners. In the initial stages of this research, we assumed that the act of holding the hand units and being asked to evaluate would affect cognitive learning. Looking at the global results, however, it seems that simply holding the hand units and being asked to evaluate is not what differentiates the obtrusiveness of the PEAC system methodology from traditional techniques, nor what defines the task demand of evaluative viewing. Study 1 PEAC users hardly ever used their hand units during the program, and not surprisingly, no differences emerged.

This suggests that the critical task demand seems to be defined, not by the electronic tools themselves, but by the evaluation question (i.e., the overt request of directed attention). The question for Study 1 appears to have been either too complicated or too general for viewers to keep in mind while simultaneously watching the program. The question had no effect on their attention. If the learners indeed were continuously evaluating, the fact that they seldom changed their opinion renders the question inappropriate because it failed to provide information useful for formative evaluation of the program.

The simpler phrasing of the question used for Study 2, relative to the question used in Study 1, was designed to make the task easier and thus encourage more frequent responses. The question did produce an effect, suggesting that the question itself plays a crucial role in both whether people respond, and, consequently, on the obtrusiveness of the PEAC methodology. This research also suggests that cognitively oriented questions will not function well within an ETV context, as moment-by-moment changes are unlikely to emerge.

Individual Differences and Content Familiarity

Although the real-time question in Study 2 produced an increase in response rate, we suspect that the nature of the program itself led to the learners' failure to use the hand units often in both studies. As mentioned above, one of the reasons for the selection of the particular program used was that it was considered a good representation of an ETV program, offering both good entertainment and learning value. It is clear, however, that viewing of this type of ETV is not a highly affective activity. A typical program would probably not elicit the individual attitude peaks and valleys of a program dealing with a controversial subject matter or a program designed largely to entertain. In other words, this type of ETV does not usually bring out strong opinions about anything on a moment-by-moment basis.

For cases in which frequent responses are not elicited, post-presentation evaluation techniques seem the best alternative. While the use of the PEAC system does not preclude the use of post-presentation techniques (in fact, to the authors' knowledge, the system is always used in conjunction with post-viewing questionnaires), the cost and effort of using an electronic evaluation system seems worthwhile only if useful additional information is obtained. Furthermore, the level of obtrusiveness, especially at increasing levels of responding as evidenced in Study 2, may obscure the evaluation of the principle aim of the program, that is, cognitive learning. If the PEAC system is used because the ETV program is felt to have sufficient affective variability, the measure of cognitive effectiveness should probably be conducted separately (without moment-by-moment electronic measurement methodology).

Mathemagenic Activities

Based on Rothkopf s work, we conjectured that requiring learners to evaluate continuously throughout a program would direct their attention to certain aspects of the program, leaving other aspects unnoticed. The Study 1 question did

not have the positive effect on content recall that mathemagenic theory would have predicted. The question for Study 2 actually appeared to have a negative impact. The detrimental electronic evaluation effect only emerged in the short answer responses, which are more discriminating of comprehension and application levels of learning. This result suggests that use of general moment-by-moment evaluation influences an individual's deeper processing of information as opposed to simple information acquisition.

The verbatim-type questions used in the studies to assess incidental learning did not seem to be influenced by the use of electronic evaluation tools. This may be because this type of recognition item does not require conscious cognitive processing in the same manner as does short answer material involving comprehension.

Attitude and Program Evaluation Ratings

There were no differences between PEAC users and non-users for any of the attitude items. In general, viewers appear to both approve of and recommend this form of a documentary in the context in which they found themselves, that is, watching ETV in the classroom. The studies show, however, that simply liking a documentary does not necessarily imply that learning will occur, as evidenced by the mediocre achievement on several parts of the posttest. This again highlights the need for designers and producers to attend to an intentional balance of various types of objectives in the design and evaluation of instructional programs.

CONCLUSIONS

Electronic assessment tools in general, and the PEAC system in particular, have been advocated as not artificially changing the act of viewing (Millard, 1992; Radio-Quebec, 1984). These studies provide support for the claim that electronic evaluation tools may be used in a valid fashion for measuring affective variables, the domain of their traditional application. This assumes that the objectives of the program include change in attitude as a central goal. However, these results provide empirical evidence to suggest that such tools are not appropriate in the assessment of a program's cognitive learning effectiveness, and indeed that they detract from it. Given that cognitive objectives are of paramount importance to educators, this interfering effect appears to seriously question the validity of using electronic measurement tools at the same time data are sought on cognitive processing.

We are reminded that, while the data produced by these systems tend to be impressive in a technological sense, it would be a mistake to generalize their use beyond their empirically demonstrated abilities. In situations in which a program's objectives are partly concerned with affect and partly with cognitive change (i.e., learning), we suggest that an evaluation using an electronic tool be

supplemented by another evaluation technique more appropriate to assessing cognitive change, such as post-viewing questionnaires.

REFERENCES

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42,* 145-170.

Baggaley, J.P. (1987). Continual response measurement: Design and validation. *Canadian Journal of Educational Communication, 16,* 217-238.

Baggaley, J. P. (1982). Electronic analysis of communication. *Media in Education and Development, 15,* 70-73.

Bracht, G. H. (1970). Experimental factors related to aptitude-treatment interactions. *Review of Educational Research, 40,* 627-645.

Bracht, G. H., & Glass, G. V. (1968). External validity for experiments. *American Educational Research Journal, 5,* 437-474.

Cambre, M. A. (1981). Historical overview of formative evaluation of instructional media products. *Educational Communications and Technology Journal, 29(1),* 3-25.

Canadian Broadcasting Corporation (Producer). (1975). *Out of the mouths of babes* [videotape]. Montreal: Author.

Clark, R.E. (1992). How the cognitive sciences are shaping the profession. In H.D. Stolovitch & E.J. Keeps (Eds.), *Handbook of human performance technology: A comprehensive guide for analyzing and solving performance problems in organizations* (pp. 688-700). San Francisco, CA: Jossey-Bass,

Corporation for Public Broadcasting. (1981). *A comparison of three research methodologies for pilot testing new television programs.* Washington, B.C.: Author.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods.* New York: Irvington.

Dailey, J. T. (1965). *Spatial visualization test of the Dailey vocational tests.* Boston: Houghton Mifflin Company.

Daningburg, S., & Schmid, R.F. (1988). Educational television evaluation: The impact of methodology on validity and learning. *Journal of Educational Television, 14,* 177-191.

Dick, W. (1980). Formative evaluation in instructional development. *Journal of Instructional Development,* 3(3), 3-6.

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Manual for kit of reference tests for cognitive factors.* Princeton, NJ: Educational Testing Services.

Cooler, D. D. (1980). Formative evaluation strategies for major instructional development projects. Journal of Instructional Development, 3(3), 7-11.

McCain, S.J.H, Short, R.H., & Stewin, L.L. (1991). Adapting instruction to individual differences: The fading promise of ATI research. In R.H. Short, L.L. Stewin, & S.J.H. McCann (Eds.), *Educational psychology: Canadian perspectives.* Toronto, ON: Copp Clark Pitman.

Millard, W.J. (1992). A history of handsets for direct measurement of audience response. *International Journal of Public Opinion Research, 4(1),* 1-17.

Nickerson, R. B. (1979). The formative evaluation of instructional television programming using the program evaluation analysis computer (PEAC). In J.P.Baggaley *(Ed.),Experimental research in TV instruction, Vol.2* (pp. 121-125). St. John's, Newfoundland: Memorial University.

Radio-Quebec. (1985). Recherche formative pour la serie *"Arrimage",* Montreal: Author, Service de la recherche.

Radio-Quebec. (1984). *Variabilite de la reponse PEAC.* Montreal: Author, Service de la recherche.

Romiszowski, A. J. *(W81). Designing instructional systems.* London: Kogan Page.

Rosenberg, M.J. (1990). Performance technology: Working the system. *Training,* (February), 43-48.

Rothkopf, E. Z. (1966). Learning from written materials: An exploration of the control of inspection behavior by test-like events. *American Educational Research Journal, 3,* 241-249.

Rothkopf, E. Z. (1970). The concept of mathemagenic activities. *Review of Educational Research, 40,* 325-336.

Scriven, M. S. (1967). The methodology of evaluation. In R. W. Tyler, et al. *Perspectives of curriculum evaluation* (AERA monograph series on curriculum evaluation, no. 1), (pp. 39-83). Chicago: Rand McNally.

Snow, R. E. (1978). Theory and method for research on aptitude processes. *Intelligence, 2,* 225-278.

Watts, G. H., & Anderson, R. C. (1971). Effects of three types of inserted questions on learning from prose. *Journal of Educational Psychology, 62,* 387-394.

Weston, C. B. (1986). Formative evaluation of instructional materials: An overview of approaches. *Canadian Journal of Educational Communication. 15,* 5-17.

ACKNOWLEDGEMENT

The first author gratefully acknowledges the support of the Social Sciences and Humanities Research Council of Canada and the Fonds pour la Formation des Chercheurs et 1'Aide a la Recherche of Quebec for initial support for this research. The authors would like to thank Robert M. Bernard for his helpful suggestions regarding the research design.

AUTHORS

Suzanne Daningburg is an Invited Senior Researcher at the Centre for Information Technology Innovation, Industry Canada, 1575 Chomedey Blvd., Laval, QC, Canada, H7V 2X2. This research was completed when Dr. Daningburg was an Assistant Professor in the Department of Education at Concordia University.

Richard F. Schmid is an associate professor in the Educational Technology Program, Department of Education, Concordia University.