Continuous Response Measurement and the Dr. Fox Paradigm

Robert Elie Salhany

A Thesis

in

The Department

of

Education

Presented in Partial Fulfilment of the Requirements
for the Degree of Master of Arts at
Concordia University
Montreal, Quebec, Canada

August 1993

ISBN 0-315-87282-9

ABSTRACT

Continuous Response Measurement
and the Dr. Fox Paradigm

Robert Elie Salhany

This study examined the validity of the Program Evaluation Analysis Computer (PEAC) with regard to the predictability of the collected data. Using the Dr. Fox paradigm, the PEAC system was used by students to evaluated segments of a video tape with regard to an actor/lecturers degrees of content and expressiveness.

Two groups of 19 students, one group evaluating expressiveness and the other evaluating content, were used as subjects. To investigate this effect a between group multivariate repeated measures analysis of variance (Manova) was performed. To further explore the hypothesis two separate 2 by 2 multivariate repeated measures analysis of variance were also performed.

The omnibus multivariate repeated measures analysis of variance showed a significant interaction effect on group by condition. This interaction indicates that expressiveness raters are quite capable of recognizing

when expressiveness is high or low, and that content raters are somehow influenced by the type of segment they are rating. Further analysis suggested that even though content was high in both instances, students rated it higher when expressiveness was also high, and lower when expressiveness was held low.

Acknowledgments

The completion of a thesis requires inspiration, perspiration, encouragement, unselfish help, and the belief that not only is the end near, but it's all worth it.

I would like to thank the following individuals:

Jon Baggaley for encouraging and helping me to pursue an idea that would have gone unnoticed;

Karen Cayrol who's knowledge and help was essential in its completion;

Mary Gardiner for her relentless encouragement and constant unselfish inspiration, when all seemed lost;

Albert, Connie, and Sandra Salhany for just being proud, and that's all it took;

Concordia University for giving me new life long friends; Julie, Eric, Steve, Ann, and many more; and

Alice and Edward Salhany who were and continue to be, the driving force in my life by giving me the confidence to persist.

Dedicated to

Edward Lutfy Salhany

and

Alice McCool Salhany

Table of Contents

List of Figures and Tables viii

1. Rationale 1

2. Literature Review 6

Microteaching 6

Formative Evaluation and the PEAC System . . 15

Dr. Fox Paradigm 26

Summary 28

3. Method 30

Subjects 30

Materials 30

Procedure and Design 31

Data Collection 34

4. Results 35

Content Experts 48

5. Discussion 51

References 56

Appendix A 61

List of Figures and Tables

Figure 1. Raters evaluations 39

Figure 2. Combined Means 41

Figure 3. Content Raters Ratings 43

Figure 4. Individual repeated segments ratings 47

Figure 5. Mean rating of content experts 49

Figure 6. Ratings of individual content experts . . . 50

Table 1. A double repeated measures design 33

Table 2. Cumulative timing points 34

Table 3. Means and Standard Deviations 38

Table 4. Combined means of expressiveness 41

Table 5. Combined means of content upon expressiveness 43

Table 6. Univariate Effects of repeated measures . . . 46

Rationale

As noted by Baggaley (1986), there is much evidence to support the finding that the application of continual response measurement techniques is increasing the precision of social science research in areas of learning and instruction. Using second-by-second continual response devices, educational technologists can gain immediate feedback of results on the impact of their educational productions. This feedback is especially useful during the formative evaluation stage of design and development.

Continual response technology is not as novel an idea as one would imagine; mechanical systems measuring the relationship between psychological response and heart beat rate can be traced back as far as 1870. A french physiologist, Jean Etienne Marey, developed the mechanical system.

Audience response technologies can be traced back to the 1930's. During the last 50 years, social scientists have been increasingly interested in the influence of mass media on society. In the 1940s and 1950s there were studies which examined reasons people read newspapers, the influence of radio soap operas on

housewives, how comics affect children, reasons for listening to classical music radio stations (Lazarfeld & Staton in Millard, 1992) and the influence of motion pictures on traditional value systems (Millard, 1955).

Computer-based continual response technologies emerged in the 1970's, when a number of models became available for use. Systems which provide an immediate summary of responses to a lecture or educational production are now common in many universities.

However, despite the apparent benefits of such technologies, the approach has had only a moderate impact on the technology of instructional evaluation and design. Questions remain as to the reliability and validity of continual response data - depending on the subjects ability to deal with the complexity of the response task and because of the requirement for criterion-referenced data, the reliability of the interpretation of results can be tenuous. The data collections and analysis techniques used in audience response measurement must be carefully reviewed.

The evaluation of teaching effectiveness is an area in which continual response measurement has been used. Studies by Centra, Leventhal, Perry, Abrami, cited in

Marsh (1985) reveal that the evaluation of teacher effectiveness serves several purposes. These studies provide substantial feedback to lecturers about the effectiveness of their teaching and provide students with information on which to base their course selections. Also, school administrators make use of the information during talks on tenure/promotion decisions.

Thus, the evaluation/revision process has been used increasingly to judge products or productions as to their educational usefulness. Whether an audience learns from the production is another question. The Dr. Fox phenomenon is described as the influence of an instructor's personality or expressiveness upon student evaluations so that ratings of enthusiastic presentations containing a lot of information differ little from ratings of enthusiastic presentations containing little or no information. The student is actually seduced into believing he/she learned. Studies conducted by Ware and Williams (1975, 1976a) have shown that, "... a) an enthusiastic presentation manner results in greater student learning when initial motivation to learn is low, b) differences in information-giving produce corresponding differences in student

learning levels, c) student/faculty ratings are valid in relation to information-giving and group-learning when presentations are not given in an enthusiastic manner, and, d) the latter is not true when faculty presentations are given enthusiastically."

The Dr. Fox Effect is a phenomenon which manipulates expressiveness which may in turn manipulate the validity of student ratings of lecturers as well as student achievement.

The purpose of this present study was to test for the Dr. Fox effect using the PEAC system as a method of collecting feedback data in a microteaching setting. The microteaching setting was chosen for the study as its goal is to identify positive and negative teaching behaviour so that the trainee may improve upon his teaching effectiveness. As such, this controlled environment allows for a comprehensive feedback dimension which plays a crucial role in the assessment stage of microteaching. A videotape was chosen as media as new research in technological configurations in teacher training suggests. "Another direction [would be] to investigate the effectiveness of new video configurations in teacher training programs..." as well

as to study how video technology is presently used in teacher training in order to raise questions which may be more appropriate than those driving current research. In order to answer these questions, the question as to whether the data gathered by the PEAC system has predictive value must be answered.

Literature Review

Microteaching

Prior to the 1960's, teacher training programmes contained many hours of observation and teaching before the candidate was assumed to be capable in the classroom setting. Hence, the development of microteaching at Stanford University during the early 1960's represented a significant change from traditional teacher training programmes (Mcknight, 1980).

Microteaching assumes the position that "...teaching is a complex set of capabilities which students in training and perhaps teachers in service have not perfected" (Trott, 1977, p. 9). A controlled environment is therefore desirable whereby class size, lesson length, and teaching complexity are scaled down. Teaching may be developed more effectively when each teaching skill is identified and practiced on its own later to be combined with other skills when mastered (Trott, 1977). The original Stanford model encompassed a six stage cycle: plan, teach, critique, replan, reteach, reobserve. The critique sessions provided feedback to the trainee through a tutor (supervisor), pupils, trainee's peers, or by various types of rating

form. Audio and video devices were also used in providing feedback to the trainee (Geddes, 1979). McKnight (1980) described a typical microteaching programme as such:

> ...after a technical skill of teaching has been described for the trainee through a videotape of a master teacher modelling the skill, the trainee's subsequent brief teaching performance is videotaped and then reviewed under various supervisory and other feedback contingencies. The videotape recordings that were made of the lesson are played back at this time. Following the critique, the trainee revises his lesson and teaches it again, usually to a different group of pupils. The second teaching session is also followed by a critique (p. 214).

The teach, assess, reteach cycle is viewed as the basic element within the microteaching environment by both Trott (1977) and McKnight (1980). Geddes (1979) characterises microteaching as having two essential features, "... a simplification of the teaching situation and the provision of feedback" (p. 3).

In essence the goal of microteaching is to identify positive and negative teaching behaviour so that the trainee may improve upon his teaching effectiveness. Cooper and Allen, cited in Griffiths (1977a) assert that in terms of changing the behaviours of trainees the feedback dimension is probably the one that is most

crucial to microteaching. The feedback mechanism is that which helps to indicate the positive and negative aspects of the trainees' teaching behaviour, and is therefore one of the main causes of concern within the microteaching setting (Griffiths, 1977a). Wragg (1970) emphasises, and adds to this point, that the source from which feedback is obtained is of great importance since it plays a major role in the assessment phase of microteaching, with variations of effectiveness by each contrasting source. Trott (1977) and Griffiths (1977a) cite four main sources of feedback which are most widely used:

1. mechanical recordings of the lesson on video or audiotape.
2. tutor supervisor feedback during the critique phase.
3. colleagues or peers.
4. pupils who make up the microclass.

Video vs Audio Tape

In comparing audio and video feedback Gall et al indicated that both were generally equal in producing skills outlined in a minicourse titled "Individualising Instruction in Mathematics" (cited in Griffiths, 1977a).

Yet a study by Ward (1970) also found that unless otherwise viewed with a model videotape of the desired behaviour, the impact of videotape feedback in altering the use of questions by in-service teachers was ineffective, but the behaviour was changed effectively by audiotape feedback only.

Shively, Mondfrans and Reed (1970) compared four conditions of microteaching in examining feedback mode. The four treatment groups consisted of supervisor based critique on a videotape (VT group), an audiotape (AT group), a live lesson (LL group), and responses from microteaching students to the Standford Teaching Competence Appraisal Guide (STCAG; SR group). Shively et al conclude that the AT group produced the greatest amount of change but that the bases for the supervisor's critiques varied across all groups. They also explain that most of the focused skills during the experiment were verbal which would result in the AT group reviewing the bulk of the crucial information. The LL treatment group was the least effective with Shively et al explaining:

> The possibility exists that a single supervisor may have introduced some bias into the results of the experiment by praising one

> form of feedback over another or by presenting the microteaching teachers with different kinds of information in the critique sessions (p. 9).

Since the VT treatment was virtually a reliving of the teaching experience, the confounding issue of supervisor's bias may also have been present. In discussing the results of the VT treatment group, Shively et al stated that, "Much of their reaction to the videotape appeared to be centred around how they looked rather than to the critical aspects of their teaching behaviour. Thus, the attention paid to the aural information was probably less" (p. 9). It may be the lack of focus and the novelty of the video media that confounded the results of this group.

It seems highly likely and logical that those in the AT group would change their behaviour more easily since the nature of their group and the experiment may have created a more focused environment. As mentioned by the researchers, the visual information in the VT treatment may have been irrelevant (i.e. verbal skills measured) and attention-getting but it may have also confused the teachers and for that matter the

supervisors who were required to base their assessment on these tapes.

All of the above mentioned studies assessed audio and videotaped feedback in measuring their effects upon verbal skills. Although audiotape is seemingly equal to or sometimes more effective than videotape, it is doubtful that this would be true for nonverbal skills. The effect of viewing both non-verbal and verbal skills, their relationship upon each other and therefore upon the trainee/supervisor relationship may be profound and far more complex than these studies could have indicated.

Video Tape

In a review of literature relating to the use of videotape technology in teacher training, Frager (1985) describes a study by Anderson, Frager, and Boling (1972) comparing videotape demonstrations and role-playing simulations. They conclude that

> ...videotaped demonstrations were found to be superior in eliciting competent instructional implementation...[of the same technique, and that the videotape]...focused viewer attention on the important teaching behaviours as they occurred (p. 21).

Legge and Asper (1972) studied the evaluations by preservice teachers with regard to their viewing of a film of a teacher's performance. They concluded that teachers who made and viewed videotapes of their own teaching rated the teaching performance significantly closer to master teachers than those who did not.

In a study of split-screen videotaping cited by Frager (1985), Mertz found teachers in the videotape treatment changed their verbal and nonverbal behaviour significantly positive with the effect maintained over a 4 month period.

Ajayi-Dopemu and Talabi (1986) studied the use of videotape recordings as an effective method of teacher education. Students were assigned to one of two microteaching groups. The first group was allowed to practice teaching with the aid of videotape equipment while the second without the aid of videotape equipment. It was discovered that the first group demonstrated a significant improvement in the mastery of the specified teaching skills.

Colleagues or Peers

Feedback from colleagues or peers has not seen as much success as mentioned by Griffiths (1977a) in citing

a study by McIntyre who found students working with peers to express a low moral which was also "reflected particularly strongly in a weaker commitment to teaching careers" (p. 21).

Cripwell (1979) in describing microteaching, used at the University of London Institute of Education, points to the dangers of peer role-playing as part of the microteaching setting, particular to the trainee's understanding of the artificial nature of microteaching:

> They (peers) seem to find the additional artificiality of role-playing in a simulation exercise difficult to accept. Unless the role-playing is carefully structured the 'pupils' tend to concentrate on their performances rather than on what their 'teacher' is doing. This can become most frustrating for both teacher and taught. The fact that peers who are as competent as the teacher does not help either in terms of the microteaching or in terms of simulation (p. 41).

There seems to be a danger in allowing peers to judge each other while they are still in the process of mastering skills.

Students Feedback

Student feedback in changing the behaviour of teachers has also been examined in several studies. Ryan (1966) examined three methods of pupil feedback and

found no significant effects. It was also reported that 82% of the pupil responses and comments were positive, thereby generally providing the teacher with little impetus to change behaviour.

Indeed, it could be argued that the students' positive responses encouraged the teacher not to change their behaviour (p. 2089-A).

Gage, Runkel and Chatterjee (1963) not only revealed significant change in teacher behaviour through pupil feedback, but also improvement in the accuracy of teachers' perceptions with regard to pupil opinions. Shively et al (1970) found pupil ratings (STCAG forms) produced changes in teaching behaviour when administered by supervisors.

In summary the feedback mechanism within the microteaching setting has cause for concern with the inconsistent findings by previous studies (Griffiths, 1977a; Trott, 1977). While research in the use of video technology in teacher training show a positive view there are still limitations to its effectiveness (Frager, 1985).

Formative Evaluation and the PEAC System

> Evaluation implies that one is seeking to assess the value of a certain activity or product of our system (p. 366, Romiszowski, 1984).

The process of evaluation implies a comparison of an act or a product to some preconceived criteria in order to determine its worth. In general, the evaluation process can turn out to be very subjective, unless some sort of data are used as part of the criterion. Formative and summative are two forms of evaluation which occur during and after the process, respectively (Lemler; Lashley & Watson; cited in Cambre, 1981).

Summative evaluation is the summation of the results or the effect a product has had on a population. One accumulates the data at the end of an event and then concludes through preconceived criteria as to whether the objectives and goals have been achieved. This type of evaluation has no immediate affect on the structure or process of the event.

The evaluation / revision process (prior to a products completion) has often been used to judge products or productions at various points in time.

Although Michael Scriven coined this procedure as 'formative evaluation' in 1967, this form of assessment emerged in the 1930s from the field of mass communications. Its purpose is to measure effectiveness in order to improve the product by modifying the structure and/or the content (Cambre, 1981).

Formative evaluation of instructional materials has proven to be very effective in development of the final product. The use of subject matter experts, learners, colleagues, and even self-critique, have proven effective in producing favourable results (Weston, 1986).

Evaluations of educational media materials were first studied during the 1920s. Educators adapted audience response equipment to use in the classroom to test student learning, using structured multiple choice questions. "Results were available immediately for the class as a whole, as well as giving a record for individual students. Student response was sometimes measured on a continuous basis to instructional television programming" (Carpenter 1950 in Millard 1992).

One instructional film which underwent substantial evaluation was entitled "Fit to Win". The massive collection of evaluation data gathered on this film helped to set this film as a significant predecessor to the formative evaluation process.

The collection of empirical data has proven not only to be necessary nor merely does it contribute because of its quantitative nature, but vital, owing to the subjective nature of the evaluation process in general.

Electronic devices used for the gathering of such evidence have proven to be useful and predictive in the evaluations of films. The evolution and implementation of such systems for the purposes of formative evaluation have been thoroughly documented (Cambre, 1981; Gooler, 1980; Baggaley, 1987). Millard (1992) notes that electronic handsets were initially invented to allow audiences to signal on an individual basis and a continual basis as to their likes/dislikes or some other evaluative criterion. People involved in research then were provided with a record of experience over time. As Millard (1992) explains:

> The unique aspect of this types of research is that it provided a record of the experience over time which proved useful on the one hand as a kind of X-ray of the exposure process (continuous measurement of cognitive processing) and helped explain the viewers' positive or negative retrospective appraisal of the program as a whole...

Many different electronic devices have emerged over the years. A doctoral dissertation by Frank Stanton in the 1930's reviewed numerous methods in use for measuring radio listening and presented a report on the experimental placing of a recorder he designed to be installed in homes to track radio-set use. This system became the benchmark for nation-wide research on audience listening and eventually television viewing by the A.C. Neilsen Co.. At Ohio State university, Dr. Stanton worked on the polygraph, "...and it was technology for simultaneously charting different types of polygraph response that he later used in designing the first electronic audience response system" (Millard, p.3 1992).

Other systems known to have existed are well documented by Millard (1992), in his paper, entitled "A History of Handsets for Direct Measurement". The Lazarsfeld-Stanton instrumentation is among one of the earliest

handsets. This device consists of two cylindrical handsets about five inches long ard each cylinder has either a red push button at the end or a green button at the end. The idea was to have each member of the audience hold one cylinder in each hand, with his thumb in a position to press down on the button. When the audience member felt he liked the program he/she was hearing, he/she was to press on the green button and was instructed to keep pressing the green button as long as the program was to his/her liking. In the event the audience member was not liking a particular portion of the program, they were instructed to press down the red button and to continue holding down the red button as long as they disliked the program. If the audience member was feeling indifferent, he was to push neither the green button, nor the red button. Registered responses were gathered from each individual member of the audience and charts of results over time were developed from sample sizes ranging from 75 to 100 (multiple sessions). Questionnaires were then filled out and a group discussion followed the session. CBS now can accommodate up to 30 viewers at one research showing (Millard,1992).

Another handset system developed is known as the Hopkins Televote Machine, which was used primarily to gather viewer information for the motion picture industry. This system was slightly different from the Lazarfeld-Stranton system in that the range of possible reactions was higher.

> Using an audience response system called the 'Hopkins Televote Machine" viewers signalled on a continuous basis whether they liked what there were watching very much, or simply like it, or felt neutral, or found it dull, or very dull (5 point scale). Responses were cumulated and recorded as a single continuous line on a graph, called the motion picture profile.
> Millard, 1992, p.4

This system boasted a dial handset with five positions to choose from and was later improved in order to provide researchers with refined information concerning the distribution of responses over categories and the linear comparison of reaction by male and female viewers.

In 1954, research conducted by McCann-Erickson on responses gathered from the Lazarfeld-Stanton system found that, "...sometimes the 'indifferent' measure reflected predominantly favourable opinion, but not sufficiently strongly to warrant holding down the green button, whereas on other programming non-response turned

out to represent mildly negative feelings, which respondents did not feel sufficiently strongly about to indicate by holding down the red button" (Millard, 1992, p.6). The device was subsequently improved to capture a four point scale. The five point scale was specifically rejected on the basis that the researchers concluded that the central point on the scale was essentially a no answer. "In effect it constituted a 'don't know' response, with the problems this entails for statistical treatment, especially with small samples of viewers. ...A five-point scale would continue the problem of the 3-point scale of the Lazarfeld-Stanton system, one step removed" (Millard, 1992, p.7). The system described here, with new unique characteristics such as a single 4-way signal switch and an automatic reminder light is known as Televac.

The PEAC system, (Program Evaluation Analysis Computer) an electronic moment-by-moment data capturing device, was developed in the 1970s. The system consists of battery charged hand-held-units (slightly larger than a calculator) whereby the operator presses one of any number of 2 - 16 keys (i.e. dependent upon the requirements of the study) in order to register a

response. Reactions are registered in real-time as a film or video is being shown, hence the subject is prompted only by ones own judgement. The results have proven to be significantly useful in formative and summative evaluation studies where the PEAC system has been employed (Baggaley, 1987, 1986a, 1986b, 1986c, 1985; Baggaley & Smith, 1982; CPB, 1981).

One such study ,conducted by Baggaley (1982), in which the PEAC system was used to track the continual responses was a study on the impact of smoking prevention films. The Canadian Cancer Society commissioned an evaluation of four films used to inform the public on the dangers of smoking. The Cancer Society officials were particularly interested in finding out the films' impact upon less-educated viewers. The 'functionally illiterate' represent a quarter of the Canadian population. The sample of viewers was primarily urban and rural communities of Newfoundland and Labrador.

Using the PEAC system information as to viewer reaction was obtained; information as to the attitudes toward the films and the films affect on prior attitudes about smoking.

The results of this study confirmed that adult illiterate persons (reading grade 8 or lower) gain little if no benefit from the films, nor from the print materials accompanying the films. The results also suggested that smoking prevention films in current use serve to reinforce non-smokers in their distaste for smoking, though they are generally received by smokers with defensiveness and hostility. Smokers are willing to consider practical guidelines for smoking cessation, being inclined to respond most positively to the films during segments when practical quitting tips are given. Viewers responded negatively to any suggestion by the films that quitting will make them appear socially eccentric. "A successful formula for films encouraging smoking cessation would be to increase viewers pessimism regarding the dangers of smoking, while reducing pessimism regarding cancer prevention and cure" (Baggaley, 1982, p.35).

The previous study gives credence to the importance of both formative and summative evaluation and that the information gathered in this manner can be extremely valuable to the film maker. In the Cancer Society study, the summative implications concern the

distribution and utilization of the film (i.e. little or no benefit to the functionally illiterate). The formative implications concern the effects of specific production techniques which affect the audience in their perceived likes or dislikes (i.e. the perception that the film suggests that quitting will make them appear socially eccentric).

Other studies using the PEAC system, conducted by Baggaley reveal evidence that criterion-referencing using between group comparisons show that interesting results may emerge from one sub-group which contrasts that of another sub-group. One example is the work done by Baggaley (1985) on preschool children and their response to TV puppet characters. The results showed a distinct difference in the way the boys reacted to certain characters and the way the girls responded to certain characters. Preferences were definitely different for the boys and for the girls. No such difference was found based on the cultural background of the preschoolers.

Handsets have proven to provide intriguing results given they allow for moment by moment continuous response measurement. There are generally three types

of continuous response data hardware available, all with varying characteristics which affects non-sampling error in response measurement. The design of the handset itself can affect the quality of the responses received in different ways. One is the handset featuring pushbuttons, which, "...as in the PEAC system, are actually less susceptible to psychometric error than other technologies demanding responses on an analogue dial" (Baggaley, 1987, p.228). Another is the dial audience response input device. This device is seen by some researchers to be less effective than the push button model;

> Dial based systems allow the subjects to set their responses wherever they choose with the available range. The manufacturers of dial-based systems commonly suggest that this is an attractive feature of their technology. However, greater freedom of response and potentially infinite response scale do not ultimately yield more reliable measures of psychological impact, for they are subject to constant over and under-shooting errors known as habituation and anticipation bias respectively. Psychometric error of this type is minimized when the response task is button-based, and the fixed psychological meaning of each response on the scale is clear to the respondent (Baggaley, 1987, p.228).

The third type of handset is what is known as the slider. Audiences can position the indicator at any marked point between two extremes of an extremely

positive or extremely negative opinion, usually using a scale of 1-10 (Millard, 1992).

Dr. Fox Paradigm

The evaluation of teaching effectiveness by students has been commonly collected at most North American universities and approved by most of the university staff and students (Centra; Leventhal, Perry, Abrami, Turcotte, & Kane; cited in Marsh, 1985). As stated by Marsh (1985), the purposes of these evaluations are variously to provide the following:

(a) diagnostic feedback to faculty about the effectiveness of their teaching;

(b) a measure of teaching effectiveness to be used in tenure/promotion decisions;

(c) information for students to use in the selection of courses and instructors; and

(d) an outcome on a process description for research or teaching (pp. 707).

The Dr. Fox phenomenon is described as the influence of the instructor's personality or expressiveness upon student evaluations of

college/university teaching (Abrami, Leventhal, & Perry, 1982; Marsh, 1982).

The question as to whether an intelligent audience would be able to recognize the deficiency in content, while being charmed into the illusion of having learned, was first asked by Naftulin, Ware, and Donnelly (cited in Abrami, Leventhal & Perry, 1982; Marsh, 1985). In their study, they found that students could be seduced into the illusion of having learned through an expressive entertaining lecture. Along with other criticisms of the study, learning was not measured.

In 1975, Ware and Williams (cited in Williams & Ware, 1977) produced a follow up experiment whereby learning was included as a measured variable. It was found that a high expressive instructor produced higher ratings than the low expressive instructor. In an attempt to simulate a true class setting, another experiment was conducted, exposing students to a follow up lecture (2 consecutive lectures) on a different theme. Again, high expressiveness produced both higher ratings and achievement than did the low expressive group (Williams & Ware, 1977).

A meta-analysis of all known Dr. Fox studies was conducted by Abrami, Leventhal, and Perry (1982). They concluded that "...instructor expressiveness had a substantial impact on summary and global student ratings..." (p. 454) and the difference was more than a standard deviation.

Essentially, students' evaluations of lecturers is a form of summative evaluation. The Dr. Fox experiments were examinations of the validity of these evaluations.

Summary

New research in technological configurations in the teacher training are suggested by Frager: "Another [research] direction might be to investigate the effectiveness of new video configurations in teacher training programs, such as computer-video interface, videodisc, information storage, and cable networking. A third possibility is for naturistic research, to study how video technology is actually used in teacher training programs (as opposed, possibly, to how people say it is used) and to raise questions which may be more appropriate than those driving current research" (p. 22).

It was therefore proposed that the use of an electronic device such as the PEAC system along with videotape, be used as a feedback mechanism in the microteaching setting. Such a system may help to focus the trainee, supervisor, and peers on skills that may otherwise go undetected. Before researching such a question, one assumes the predictive qualities of the PEAC system. Such an assumption could prove to be erroneous and misleading to those involved in teacher training. Hence the question of validity of the PEAC system must be answered first.

The Study

Therefore, the purpose of this research was to answer this question. Does the information retrieved by the PEAC system have predictive value? In other words, can a trend be seen from the collected data?

The Dr. Fox paradigm was used as a model in a videotape lecture (high/low segments of expressiveness x content) with subjects using the PEAC system in order to differentiate between high/low expressiveness and high/low content areas. Since the instructions for the evaluations of both expressive and content areas were varied and only one variable may be measured by the PEAC

system, two groups of students would be assigned to either the expressive or the content evaluation.

In view of the previous findings, it was hypothesised that there would be a significant percieved difference between low and high expressiveness in the video. It was simularly hypothesised that there would be no differences between perceptions of low and high content.

Method

Subjects

The sample consisted of 38 post secondary students. Participants were drawn as whole classes at the CEGEP level. They were sampled from areas of study that have more interest in the subject matter of the film, such as the humanities, communications, and the social sciences, than other departments at the CEGEP level. Three content experts were asked view and post their opinions of the film.

Materials

A 40 minute video-taped lecture on the *Effects of Television Upon Children* was shown to the participants. The lecturer was a subject-matter expert in this field as well as a professional actor.

The PEAC system, with its hand-held computerized data collection units, was used to retrieve the moment-by-moment reactions of the participants during the viewing of the film.

A questionnaire of 15 items was used to create a repeated measures design of 4 treatment conditions. Each item (i.e., question) represented 1 of 4 conditions, hence 3 different questions representing the same condition but occurring at different times during the video tape.

Procedure and Design

Two classes of 19 subjects each (n=38) were asked to view the film and evaluate the lecturer's perform nce. The first class evaluated the content, while the second class evaluate the expressiveness of the lecturer. The same forty minute videotape was administered to each of the classes. Three content experts (teachers in Concordia's Communication Studies Department) were asked to validate the content aspect of the video and their results were compared to the intended segments (i.e., treatment conditions) of the video tape. This was not meant as a statistical comparison but rather as a cursory view with regard to

intended high or low content areas of the lecturers performance.

The lecturer recited, item by item with its set of multiple choice answers and allowing some time for the viewing audience to contemplate a correct answer, a recently released survey on the *Effects of Television Upon Children* . Each questionnaire item was used by the lecturer to create a treatment condition (eg. high expressiveness X low content). After each item, the surveyed audiences' perceptions were released by the lecturer followed by the correct answer and an explanation of the latter.

During the performance of the lecture, the professor varied the content and expressiveness to the degree of creating four (4) treatment groups of high/low content x high/low expressiveness. The set of 4 treatment groups were repeated 3 times during the video (repeated measures design). Hence, the reoccurrence each condition three (3) times for each rating group during the film (see table 1).

Table 1. A double repeated measures design of occurrences and treatment conditions.

	Treatment Conditions			
Rating Groups	HE/HC	HE/LC	LE/LC	LE/HC
Expressiveness	Occur 1 Occur 2 Occur 3	Occur 1 Occur 2 Occur 3	Occur 1 Occur 2 Occur 3	Occur 1 Occur2 Occur 3
Content	Occur 1 Occur 2 Occur 3	Occur 1 Occur 2 Occur 3	Occur 1 Occur 2 Occur 3	Occur 1 Occur 2 Occur 3

Content was defined as the oral presentation with regard to the clarity of the presenter's answers to each item from the survey. *High content* was accurate and clearly defined factual information while *low content* was defined as ambiguous, unclear, and confusing information.

Expressiveness was defined as the dynamic visual appeal of the lecturer. *High expressiveness* was defined as an individual with a dynamic personality with more physical movement (eg. use of hands), humorous anecdotes, a voice that varied in tone (intonation) and projection. *Low expressive* features were: little body movement, monotone voice while still projecting, no humour injected.

The moment-by-moment evaluation data from each group was collected with the use of the PEAC system. Participants were asked to use a four point likert scale with buttons lettered from 'A' to 'D', sequentially, representing the following perceptions with regard to either content or expressiveness - good, fairly good, not very good, poor.

Data Collection

The hand-held-units from the PEAC system sampled each subject's selection of a button (lettered A to D) every 10 seconds which represented one timing point. Each video segment represented a treatment condition, (i.e. a questionnaire item) and varied in its allotted time frame mainly because of the differences between questionnaire items with regard to their content. Table 2 demonstrates the cumulative timing points of the 4 conditions in this repeated measures design.

Table 2. Cumulative timing points per treatment condition

	High Expressiveness	Low Expressiveness
High Content	52	37
Low Content	58	23

Fifteen questionnaire items had been performed by the lecturer. In order to create the repeated measures design of 4 conditions, data for the last 3 questionnaire items (i.e. treatment conditions) were removed from the analysis.

Cumulative data for each segment per subject was weighted in order to create a segment mean for each subject. The following is an example of a subject who pressed the following buttons (A to D; see previous section) a specified number of times during the first segment of the video tape: A=0, B=11, C=6, D=0. The following formula was used to create this subjects segment mean ((D*0)+(C*1)+(B*2)+(A*3))/17 (denominator is the number of timing points for the segment). Twelve segment means were calculated per subject (i.e. the occurrence of 4 conditions repeated 3 times each).

Results

It was hypothesised that students rating content are influenced to believe that when a teacher is expressive he also teaches content well and the reverse, if the teacher shows low expressiveness students are led to believe that content is low as well. This is what is called the Dr. Fox paradigm. However, students prompted

to rate only expressiveness are not bound to these effects, that is they are not influenced by the quality of the content. Further, the Dr. Fox effect influences the actual learning outcome in the same manner, which is however not the focus of this thesis.

To investigate these effects a between group multivariate repeated measures analysis of variance (Manova) was performed. To further explore the hypothesis two separate 2 by 2 multivariate repeated measures analysis of variance were also performed. The between group variable had two levels, content raters and expressiveness raters. Four types of segments were constructed: high content/high expressiveness; high content/low expressiveness; low content/high expressiveness; low content/low expressiveness. These conditions constitute the repeated measure, here called conditions, which also has two levels, high versus low content in the first analysis, and high versus low expressiveness in the second analysis. Each condition occurred 3 times which is the multivariate part of the analysis. An overall mean change over time is constructed from the three observation occasions (Manova, SPSS, v.4.3; 1991).

Table 3 summarises the means and standard deviation according to conditions by group and by order of occurrence (i.e., obs.1, obs2, etc.)

The omnibus multivariate repeated measures analysis of variance showed a significant interaction effect on group by condition, (the hotellings T^2 value is a multivariate t- value, that is the combined effect of the three occurrences over time) $HT^2(9,28) = 2.97$, $p < .01$. This interaction indicates that expressiveness raters are quite capable of recognizing when expressiveness is high or low, and that content raters are somehow influenced by the type of segment they rating. Figure 1 below is a graphical representation of this interaction, using the combined mean, that is the three occasions within each condition is combined to one mean for each group.

Table 3. Means and standard deviations (in parenthesis) for type of segment and rating groups

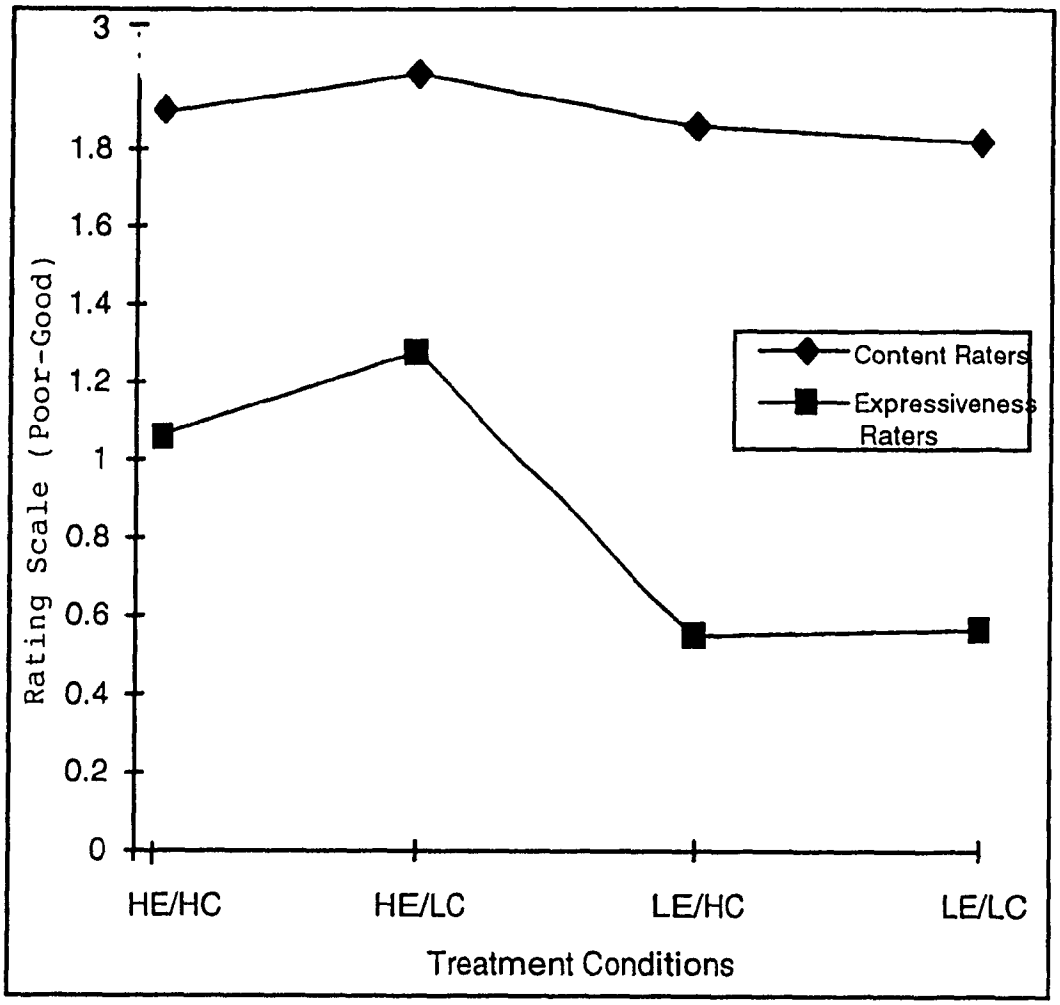
High Expressiveness Segments

Rating Groups		***High Content***	
	Obs.1	**Obs.2**	**Obs.3**
Content	2.026 (.725)	1.940 (.891)	1.713 (.956)
Expressiveness	1.233 (.781)	0.970 (.650)	0.968 (.681)
		Low Content	
	Obs.1	**Obs.2**	**Obs.3**
Content	2.065 (.352)	2.141 (.779)	1.769 (.792)
Expressiveness	1.499 (.609)	1.504 (.928)	0.818 (.690)

Low Expressiveness Segments

Rating Groups		**High Content**	
	Obs.1	**Obs.2**	**Obs.3**
Content	2.000 (.791)	1.653 (.968)	1.901 (.912)
Expressiveness	0.591 (.674)	0.557 (.621)	0.507 (.538)
		Low Content	
	Obs.1	**Obs.2**	**Obs.3**
Content	2.083 (.887)	1.800 (.857)	1.553 (.878)
Expressiveness	0.496 (.612)	0.553 (.561)	0.640 (.537)

Figure 1. Raters evaluations across each treatment condition.

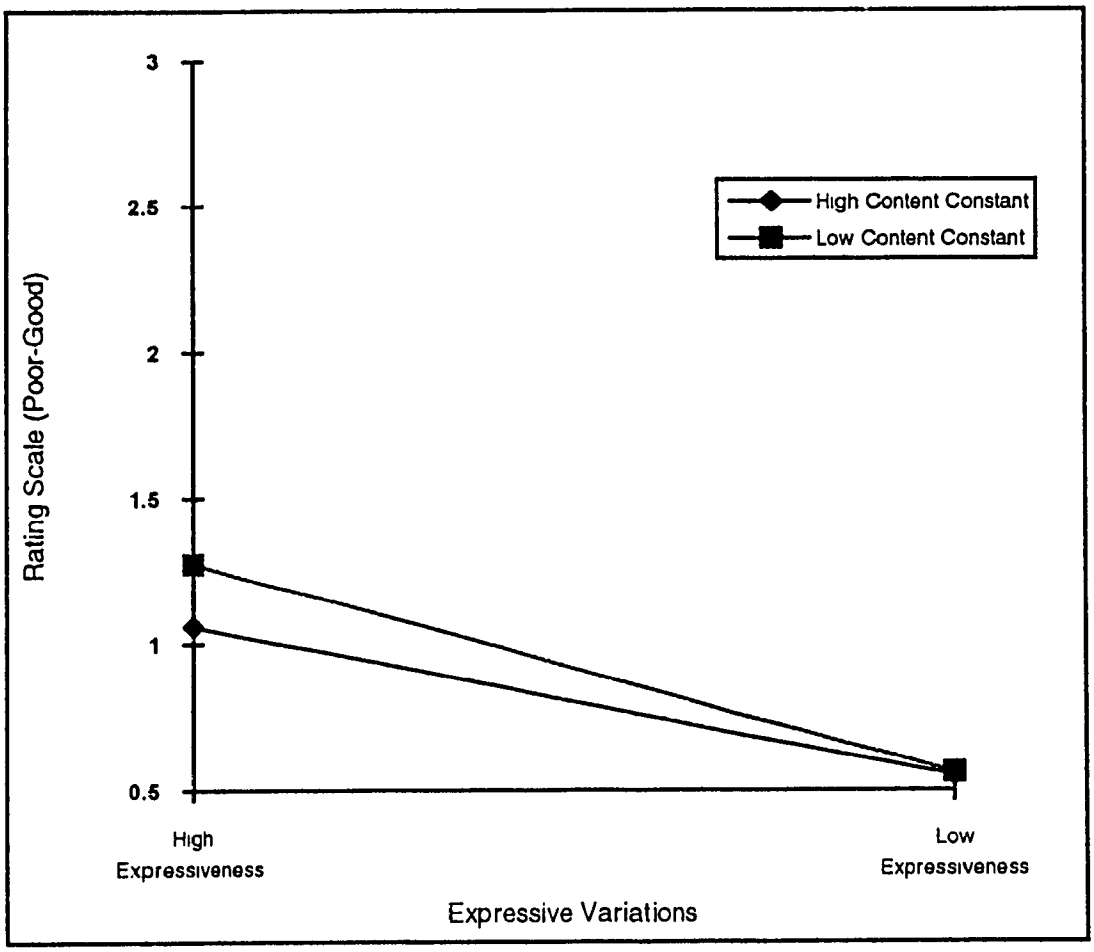


To further explore this interaction, it was necessary to establish that students prompted to rate Expressiveness were able to do so. Holding content high and looking at whether they were capable of recognizing when expressiveness was high versus low, investigation of the means showed that there was a significant difference in favour of high expressiveness $F(1,36) = 18.50$, $p<.01$. The same held true when content was held low, $F(1,36) = 35.36$, $p<.01$. The combined means are displayed in table 4. Figure 2 shows this effect graphically.

Table 4. Combined means of the effect of expressiveness upon content

Constant	Segments: High Expressiveness	Segments: Low Expressiveness
High Content	1.057	.551
Low Content	1.274	.563

Figure 2. Combined means effect of expressiveness raters.



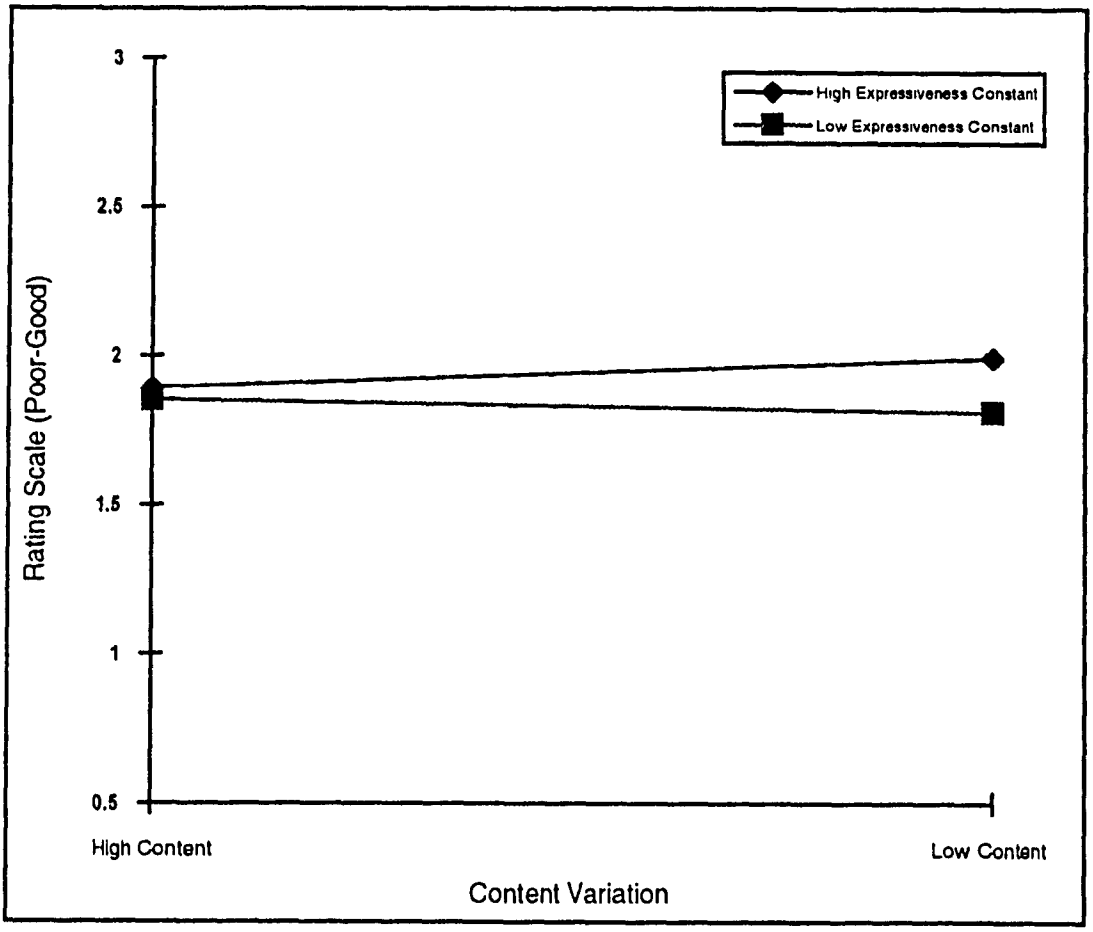
Content Raters

Table 5 shows the combined raw score ratings for the content group serving as the basis for the second analysis, however, Manova constructs a mean change over time for these values, which is why these values looks misleading. These are also represented graphically in figure 3.

Table 5. Combined means of the effect of content upon expressiveness.

Constant	Segments: High Content	Segments: Low Content
High Expressiveness	1.893	1.991
Low Expressiveness	1.851	1.812

Figure 3. Graphical representation of the content raters ratings.



Expressiveness held constant

When expressiveness was held high, that is looking at the content group's ratings on high versus low content, no significant differences were found ($p > .05$), nor when it was held low. This non-significant finding appear to indicate that performer's / teacher's expressiveness does confound the quality of the content being communicated.

Content held constant

Investigating the hypothesis that expressiveness would influence the content raters to rate higher when expressiveness was held high, than when expressiveness was held low holding content high, was substantiated, $F(1,16) = 18.5$, $p < .01$. This suggests that even though content was high in both instances, they rated it higher when expressiveness was also high, and lower when expressiveness was held low.

Examination of this effect showed that when expressiveness was high and content low versus expressiveness low and content low, they still rated content higher in the first condition, $F(1,16) = 35.36$, $p < .01$, that is the influence of the

performer's/teacher's expressiveness was confirmed. These effects are the within group multivariate effect.

Univariate Effects: Order of occurrence

It was assumed that the attention span of an audience and their ability to correctly rate a performance in terms of both content and expressiveness may taper off with time.

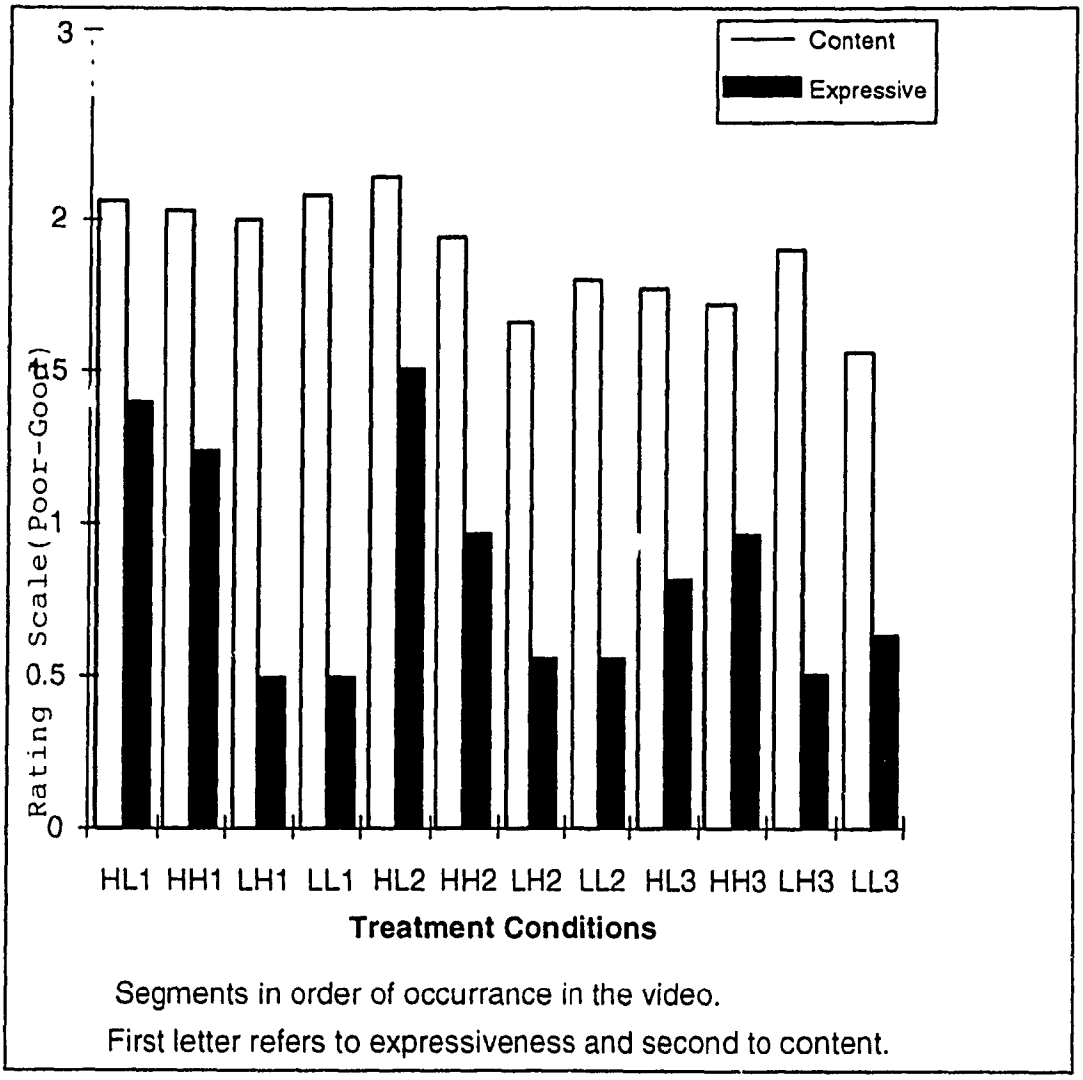
To verify this assumption early ratings (obs.1's) were looked upon multivariately but the univariate contribution in terms of condition. Table 6 shows these findings that also seem to confirm the underlying assumptions of relaxed attention towards the end of a performance in terms of rating content, however, expressiveness raters still appear to be able to correctly rate that aspect of a performance.

Table 6. Univariate effects

	Condition versus Condition	Significance $F_{(1,16)}$
Observation 1		
Content	H.Ex./H.Cont.> L.Ex./H.Cont.	2.80, p<.01
Expressiveness	H.Ex./H.Cont.> L.Ex./H.Cont.	4.21, p<.01
Observation 2		
Content	H.Ex./H.Cont.= L.Ex./H.Cont.	p > .05
Expressiveness	H.Ex./H.Cont.> L.Ex./H.Cont.	4.73, p<.01
Observation 3		
Content	H.Ex./H.Cont.= L.Ex./H.Cont.	p > .05
Expressiveness	H.Ex./H.Cont.> L.Ex./H.Cont.	3.11, p<.01

The actual order of these segments as shown in the video and the ratings by group is represented in Figure 4.

Figure 4. Ratings of each segment as occurring in the video



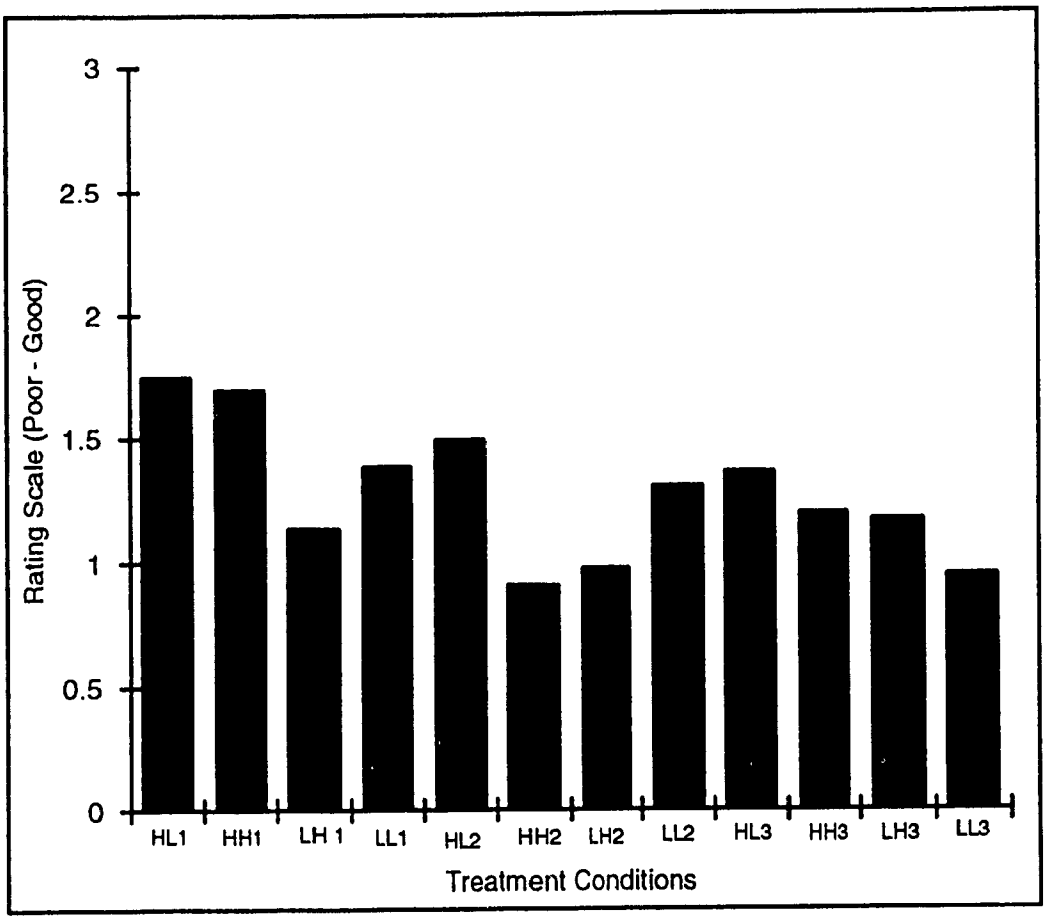
Segments in order of occurrance in the video.

First letter refers to expressiveness and second to content.

Content Experts

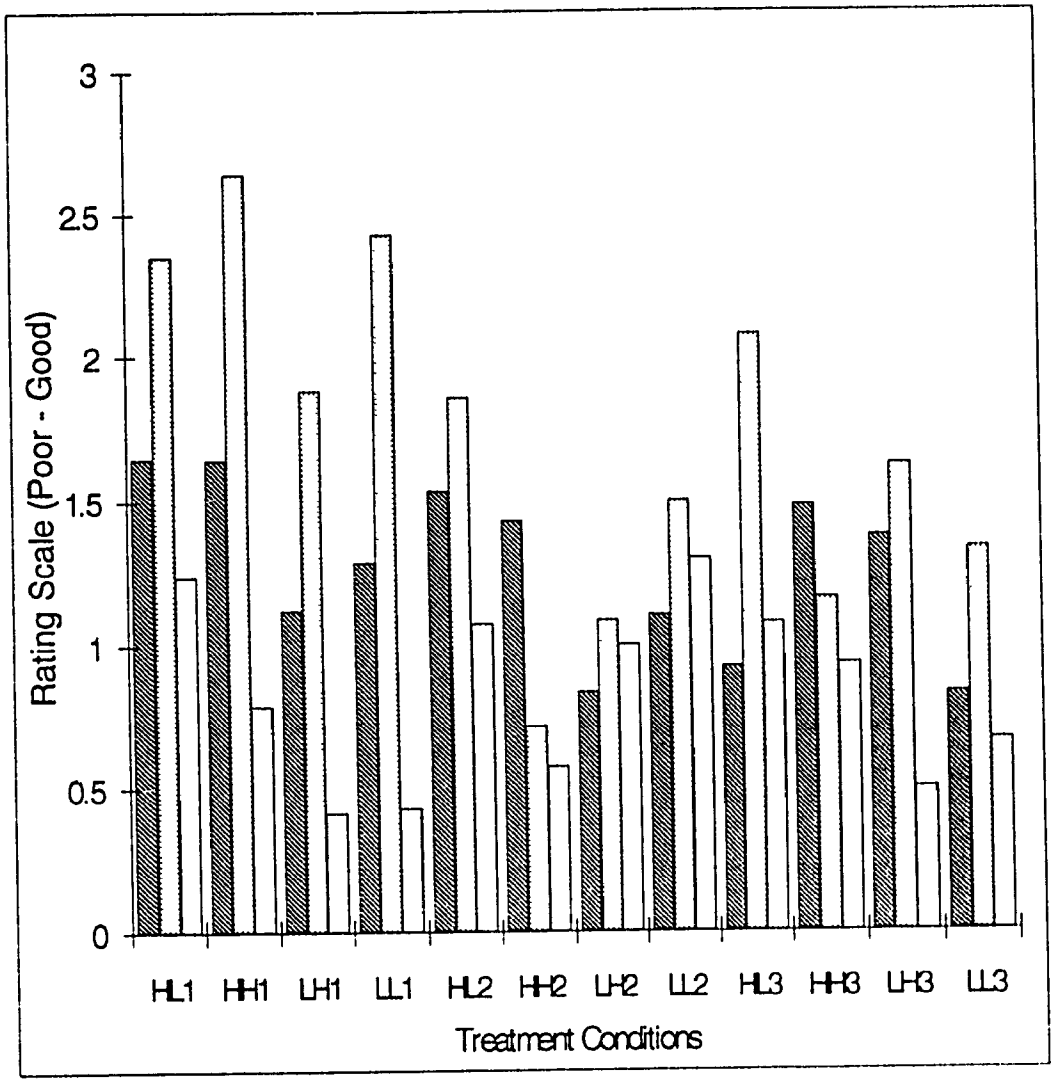
A cursory examination of the expert data reveals in generally, an inconsistency between the intended high and low content areas by the video as perceived by these experts. Figure 5 demonstrates in a graphical form, their perceptions as a group, of each segment. When further broken down as individuals, the discrepancies between each other seams to vary greatly. Figure 6 shows these variation graphically.

Figure 5. Mean ratings of content experts for each video segment.



Segments are in order of occurrance in the video
First letter refers to expressiveness and second to content

Figure 6. Ratings of individual content experts per video segment.



Discussion

It was hypothesised that there would be a significant percieved difference between low and high expressiveness in the video. It was similarly hypothesised that there would be no difference between perceptions of low and high content. The Dr. Fox paradigm was used as a comparative model in order to validate this question. The results of this study substantiate the hypothesis and therefore it can be concluded that the PEAC system does collect valid data.

The fact that the PEAC system was not only able to replicate previously researched findings which discovered distinguishable differences between high and low expressiveness, but also the system found a significant effect of expressiveness upon content raters. This leads one to believe that the system does produce valid and useful information; hence, a positive conclusion to the validation of the PEAC system.

The larger overall purpose of this research was to answer the question, "Does the information retrieved by the PEAC system have predictive value?". In other words, can a trend be seen from the collected data? The fact that raters were able to distinguish between low

and high expressiveness regardless of whether content was *held* high or *held* low through a repeated measures design demonstrates the sensitivity of the system in collecting this information.

The same did not hold true for the raters of content. They seemingly could not distinguish between confusing and inaccurate information, and clear and accurate information. One would have thought that distinguishing between clear or unclear results of each questionnaire item would have been relatively easy for content raters. However this was not the case. Perhaps the influence of expressiveness may have confused the content raters.

The fact that discrepancies occurred between content experts' ratings forces one to question what it was each expert was focusing on. Were they focusing on the same ctiteria that the students were in evaluating each segment? Or, were the content experts evaluating on the bases of whether the information given on the questionnaire was factual? Content experts may only be useful in evaluating facts of content and may not be helpful in evaluating the ambiguity or clarity of the communication of information. It may also be that they

too, were influenced to some degree by the expressive nature of the lecture.

In any case, the further analysis of student data did demonstrate the influence of expressiveness even on those rating content. Even though content was held high they (content raters) rated higher those segments of the video that were high expressive as opposed to low expressive segments. This same significant high expressive influence was produced when content was held low. This influence of expressiveness is, in essence, the main theme of what has been coined, "Educational Seduction". The belief that one is acquiring information where very little factually clear information is given is known as educational seduction.

The fact that it was easier to distinguish between a physically and vocally dynamic lecturer, as opposed to a physically stagnant and monotone lecturer, really attests to the presumption that most individuals know when they are being entertained.

The difficulty that subjects had in distinguishing clear or unclear facts leads one to believe that unless one has at least a basic grounding in the presented subject matter, it may be too difficult to judge what is

ambiguous or clear content - especially with a dynamic lecturer. Hence, the influence of expressiveness and the ramification of educational seduction.

The fact that content raters' responses tapered off with time may attest to the difficulty of not only rating content, but in keeping ones concentration and attention over an extended perios of time. This phenomina may have contributed to the overall insignificant effect of content in this study.

Summary

In conclusion, the use of a hand-held computerized device can be an asset in collecting real-time data in teacher training or any other situation where immediate feedback is critical. Its use in experimental designs may be limited only by the creativity of the experimenter and by its obvious limitations of measuring one variable at a time.

In any case, further research may be done (i.e. Dr. Fox paradigm) by using content and communication experts to measure exact segments of a video that actually contain clear knowledge and then measuring its effect upon students with regard to learning.

Further research in real-time data may be our only way of capturing the perceptions of individuals as they receive information. New technologies such as real-time computer feedback systems may be one of the most useful techniques in looking into the processes of the learners' perceptions rather than just the learning outcomes of processes.

References

Abrami, P. C., Leventhal, L., & Perry, R. P. (1982). Educational seduction. Review of Educational Research, 52, 446-464.

Ajayi-Dopemu, Y., & Talabi, J. K. (1986). The effects of videotape recording on microteaching training techniques for education students. Journal of Educational Television, 12, 39-44.

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

Baggaley, J. (1986a). Formative evaluation of educational television. Canadian Journal of Educational Communication, 15, 1, 29-43.

Baggaley, J. (1986b). Developing a television health campaign: I. Smoking prevention. Media in Education & Development, 19, 43-7.

Baggaley, J. (1986c). Developing a television health campaign: II. Skin cancer prevention. Media in Education & Development, 19, 86-90.

Baggaley, J. (1985). Design of a TV character with visual appeal for preschool children. Journal of Educational Television, 11, 41-3.

Baggaley, J., & Smith, K. (1982). Formative research in rural education. Media in Education & Development, 15, 173-6.

Cambre, M. A. (1981). Historical overview of formative evaluation of instructional media products. ECTJ, 29, 3-25.

Corporation for Public Broadcasting (1981). A comparison of three research methodologies for pilot testing new television programs. Washington, DC: office of Communication Research, CPB.

Cripwell, K. R. (1979). Microteaching in the training of teachers of english as a second of foreign language. Workshop held at the London Institute of Education. (ERIC Document Reproduction Service No. ED 224 282).

Frager, A. M. (1985). Video technology and teacher training: A research perspective. Educational Technology, 25, 20-22.

Gage, N. L., Runkel, P. J., & Chatterjee, B. B. (1963). Changing teacher behaviour through feedback from pupils: An application of equilibrium theory. In W. W. Charters Jr. & N. L. Gage (Ed.), Readings in the Social Psychology of Education (pp. 173-181).

Geddes, M. (1979). Microteaching and foreign language teacher training. Workshop held at the London Institute of Education. (ERIC Document Reproduction Service No. ED 224 280).

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

Griffiths, R. (1977a). The contribution of feedback to microteaching technique. In A. J. Trott (Ed.), Selected microteaching papers (pp. 17-24).

Legge, W. B., & Asper, L. (1972). The effect of videotaped microteaching lessons on the evaluative behaviour of pre-student-teachers. Journal of Teacher Education, 23, 363-366.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. Journal of Educational Psychology, 76, 707-754.

McKnight, P. C. (1980). Microteaching: Development from 1968 to 1978. British Journal of Teacher Education, 6, 214-227.

Millard, W.F. (1992). A history of handsets for direct measurement of audience response.

International Journal of Public Opinion Research, 4, 1-16.

Ryan, K. A. (1966). The use of students' written feedback in changing the behaviour of beginning secondary school teachers. (Doctoral dissertation, Stanford University). Dissertation Abstracts International, 27, 2089A.

Shively, J. E., Van Mondfrans, A. P., & Reed, C. L. (1970). The effect of mode of feedback in microteaching. Presented at annual meeting of AERA, Minneapolis. (ERIC Document Reproduction Service No. ED 037 391).

Trott, A. J. (1977). Microteaching - An examination. In A. J. Trott (Ed.), Selected microteaching papers (pp. 9-16).

Ward, P. M. (1970). The use of the portable videotape recorder in helping teachers self-evaluate their teaching behaviour. (ERIC Document Reproduction Service No. ED 038 365).

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

Williams, R. G., & Ware, J. E. (1977). An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings after repeated exposures to a lecture. American Educational Research Journal, 14, 449-457.

Appendix A

Questions Extracted from a study by:

The Children's Broadcast Institute, Toronto

The Children's Broadcast Institute
L'Institut de radiotélévision pour enfants

POWER OF TELEVISION
VIDEO TAPE QUESTIONNAIRE

PLEASE CIRCLE CORRECT ANSWER

1. In Canada, there is at least one TV set in (a) 50% of homes, (b) 25%, (c) 96%, (d) 75%
2. How much time do the children between ages 5 and 12 spend watching TV each day? (a) 1 to 2 hours, (b) 3 to 4 hours, (c) 5 to 6 hours
3. By the time a child has finished high schoo, he or she will have spent 11,000 hours in classrooms. How many hours will be spent watching TV during those years? (a) 2,000 hours, (b) 10,000, (c) 15,000, (d) 25,000
4. By the age of 18 an average child will have seen how many commercials? (a) 3,500, (b) 350,000, (c) 35,000, (d) 350
5. Between 4:00 p.m. and 5:00 p.m. on an average week day, half of all children in Canada are watching TV. True? False?
6. In Canada, how many national network programs are made specially for children, and broadcast in after-school hours, Monday to Friday? (a) 5, (b) 10, (c) none, (d) one
7. According to research studies carried cut in 12 countries, which of the following free time activities did the majority of subjects take part in for most of the time? (a) newspaper reading, (b) socializing, (c) television viewing, (d) studying
8. TV has caused an increase in the total amount of sleep most of us get on an average basis. True? False?
9. Which does an average Canadian child between kindergarten and high school graduation spend more time with (a) the teacher, (b) the TV set
10. Since TV, we spend less time with outdoor activities, attendance at sports and cultural events, and leisure travel. True? False?
11. Check those activities people spend less time with since the invention of TV: (a) reading, (b) conversation, (c) household tasks, (d) listening to radio, (e) play
12. Most people watch TV for: (a) information, (b) entertainment, (c) challenge, (d) instruction