Continual Response Measurement: Design and Validation

Jon Baggaley

Abstract: Computer-based measurement techniques are increasing the speed and precision of social science research methods. Using time-based polling techniques, advertising and political researchers gain rapid, second-by-second feedback concerning the impact of their media campaigns. The techniques of continual response measurement are also used in the development of educational communications, and in the 'formative evaluation' of their impact.

However, the validity and reliability of continual response data are open to question. They depend on sampling restrictions, on the complexity of the response task, and on the subjects' ability to cope with it. They require the criterion-referencing of data, and caution in the interpretation of results. The present paper discusses steps to be taken in these respects when continual response measurement is used in formative evaluation and research. Guidelines for the design of such studies are provided, examples are given typifying their deductive and inductive functions, and distinctions are made between formative evaluation and formative research on this basis.

CONTINUAL RESPONSE MEASUREMENT IN AUDIENCE RESEARCH

During 1985, an American sporting goods company announced the invention of the computerized running shoe. Following a run, the shoe is plugged into a home computer, and the runner is provided with immediate feedback regarding the distance he has covered, the time taken, and the amount of calories burned up. The concept behind the system is sound. Immediate feedback of results can be expected to increase the runner's ability to improve his skills the next time out. He no longer has to rely on

Jon Baggaley is Professor of Educational Technology at Concordia University, 1455 de Maisonneuve W., Montreal, PQ H3G 1 M8. His research and teaching interests include the psychology of communication, audience research, and the evaluation of educational television and film. The author gratefully acknowledges the support of the Canadian Cancer Society to research reported in this paper. The paper is an edited version of an audience research monograph available from the Education Department, RTV1, Johannesburg. It forms the second in a series of articles on formative research (see CJEC 1986, Vol. 15, No. 1, pp. 29-43).

intuition in order to gain the maximum return from his athletic efforts.

As indicated in the previous paper in this series, similar feedback devices have come on the market for the benefit of film and television producers. Since the development of portable microcomputing facilities in the 1980's, media producers no longer have to rely exclusively on questionnaire and interview techniques for information about their production's impact. These techniques were in any case largely unable to provide the specific information which producers require about the impact of particular production techniques. Precise moment-by-moment feedback of audience reactions to a production is now available, generated by a wide range of electronic facilities. The history of such research systems is discussed by Cambre (1981), Malik (1981), Clarke and Ellgring (1983), and Edel (1986).

Via the new systems, the audience's responses can be recorded continuously as they view a production, and fluctuations within them instantly analyzed (Baggaley, 1986a). When the continuous record of audience responses is synchronized with the production itself, the producer can inspect the momentary fluctuations in response which are associated with individual scenes and production techniques, in time to reshoot or reedit the programme for greater effect. The moment-by-moment responses may also be examined for individual differences among viewers, and insights gained from the reactions of different types of viewers to particular production techniques. Thus, reactions of viewers of different age groups and abilities to programme pacing and illustration techniques may be compared. For programme policy-makers, continual response measurement (CRM) can answer general questions concerning, for example, the reactions of different audience types to programme violence and stereotyping.

The value of research during the process of media production has been evident since the earliest days of educational film (see for example, Lashley &Watson, 1921; and Zirbes, 1924). In 1967 it was recognized formally by Scriven under the heading *formative evaluation* This term has proved most valuable for the purpose of drawing a distinction between the practical types of evaluation study conducted during the production process, and the more common forms of evaluation known as and conducted after production is completed. The latter type of study, with its tendency to expose production faults too late for producers to do anything about them, has hardly endeared the media evaluator to his or her production colleagues. In fact it has done much damage to their relationship.

The new computer-based measurement techniques promise to speed up the media evaluation process, and to create a more productive relationship between the producer and researcher. The techniques of CRM offer particular benefits. However, they are unlikely to be used widely until the data collection and analysis procedures on which they are based have been carefully reviewed. For, as the following article indicates, the reliability of continual response data are often questionable, and the validity of the results are thereby jeopardized. These problems must be carefully kept at a minimum in the design and interpretation of formative evaluation and research studies generally.

COMPARISON OF CONTINUAL RESPONSE METHODOLOGIES

Response analysis systems differ on a large number of bases: notably portability

and flexibility, speed and level of analysis, clarity of feedback, and the combination of these facilities relative to cost. In broadcasting research, of course, the common need is for multiple hand-units to record the responses of a whole audience. The *Program Evaluation Analysis Computer* (PEAC system) collects the responses of a potentially infinite number of people, via a set of remote battery-powered units. Reactions to a production may be collected simultaneously in a range of settings (e.g., viewers' homes) as well as in a central location (Nickerson, 1970; Baggaley, 1986a).

A typical system involves a series of hand-held response units via which observers' responses are collected, and transmitted to a computer for analysis. Changes in, for example, the frequency, average length and variability of behaviour can then be examined across time. To the designers of a TV or film production, of course, such feedback about its moment-by-moment impact can be irresistible. Programme segments can be adjusted or extended and camera angles altered — even during live presentations — in order to maintain and enhance programme appeal. However, the quality of such feedback is only as sophisticated as the research methods which were used to generate it; and impulsive interpretations of hastily gathered data can be highly suspect.

In educational media research, for instance, the extent to which a measure such as moment-by-moment appeal can actually predict overall learning is debatable. Similarly, little is known about the criteria by which a meaningful shift in response can be distinguished from a random one. Many media producers are rightly defensive about the introduction of continual response methods into their industry for such reasons. They suspect that audience researchers will use the methods to dictate aspects of production content on quite unjustified bases. Urgent attention must therefore be paid to the research methodology on which such systems depend, while the broadcasters are still willing to consider their benefits.

A comparison among three of the leading methodologies in North American broadcasting research was made by the Corporation for Public Broadcasting (1981). Programme pilot-test results obtained by two of the electronic methods (the PEAC system and the Percy Voxbox) were compared with those of a more conventional testing method, the discussion or focus group. The three approaches were judged in terms of a) response articulateness versus objectivity; b) sampling flexibility; c) practical benefits to programme producers; and d) long-term benefits to programme policy-makers and distributors.

The conclusions of the CPB study may be summarized as follows. While the openness of the focus-group situation usually allows discussants to be relatively flexible and uninhibited in their responses, it can also have inhibiting effects. Powerful group biases can affect the opinions expressed. The opinions of individual group members may be dominated by those of more assertive individuals. By the time the presentation is over and the discussion takes place, viewers may also have forgotten many of the critical but fleeting reactions they experienced whilst the presentation was in progress.

The availability to record one's responses to a programme simultaneously, via a hand-held response unit, can reduce these problems. Being nonverbal, responses are usually private and anonymous. Audience members have the opportunity to make a completely uninhibited assessment, and to change it as frequently as they choose. On

the other hand, an automated response task invariably restricts the range of available responses to a set of fixed options. The CPB study (1981) concluded that the most effective testing situation for the foreseeable future will probably be one featuring the electronic and focus-group methodologies simultaneously.

Certainly, the electronic techniques are the only current means whereby moment-by-moment fluctuations in audience impact can effectively be measured: one would hesitate to stop the programme every few seconds for a discussion! The imposition of a closed-ended response can be seen as a worthwhile price to pay for this extra information. On the other hand, the overall impact of a programme is unlikely to be established other than by post-test measures (e.g., questionnaire or discussion methods).

The strengths and limitations of CRM methods are indicated by the following case studies. The studies were conducted by the author between 1980 and 1984, initially at Memorial University of Newfoundland, and more recently at Concordia University, Montreal. Both universities had purchased, for their media research purposes, the Programme Evaluation Analysis Computer. The PEAC system was selected from the range of possible systems on the basis of its superior portability and flexability of operation, and its relative cost-efficiency.

SAMPLING RESTRICTIONS

Since electronic hand-units are more expensive to obtain than questionnaires or telephone calls, the samples of the population with which they can be used are usually more restricted. Unless an adequate sample can be amassed via several test sessions, the external validity of research results is likely to be restricted. The problem commonly arises in formative evaluation studies requiring rapid feedback of results to, for instance, a programme producer. It also occurs when audience reactions to a Iive, one-shot media presentation are studied, so that an immediate analysis may be obtained while the subject matter is still topical.

In November 1980, the PEAC system was used to assess public reactions to the televised debate between American President Jimmy Carter and the presidential challenger Ronald Reagan. In St. John's, Newfoundland, a panel of two dozen viewers watched the debate in their homes. As they did so, they used the portable PEAC hand-units to respond to the following question: "Who, from one moment to the next, is winning the most votes?". Three options were available to them, on buttons labelled CARTER, REAGAN and DON'T KNOW. Their continual responses were sampled at 4-second intervals. Although limited in its scope and generalizability, the study gave indications of the telling impact of nonverbal strategy in the debate, and of the speed with which the contender Reagan was able to dominate President Carter in the viewers' eyes. The study has been described in more detail in the preceding article in this series 1986a).

The rates of audience response during the first eight minutes of the Carter-Reagan debate are plotted graphically in Figure 1. The four graphs are divided into segments, according to the alternating question-and-answer format of the debate. The first four minutes is dominated by responses on the DON'T KNOW button (Fig. la) and the

CARTER button (Fig. lb). Viewers perceived the incumbent Carter as winning the most votes even before he spoke. During the fifth minute, however, votes began to accumulate for Reagan (Fig. lc), and in the subsequent course of the debate, Reagan's perceived *votability* increased dramatically, particularly in the 19th minute during his discussion of Carter's economic record. The peaks of response on the three buttons may be compared in the combined graph (Fig. 1d).

The problems of inferring overall impact from such data are obvious. Firstly, the panel of subjects used in the Newfoundland study was minimal in size, and any attempt to generalize from their responses to the larger American audience would be highly questionable. (One can sympathize with the accused in a court of law, for whom life and death depend on the reactions of a jury half this size.) If such a study is to be beyond reproach, therefore, it must clearly make use of *a representative and balanced sample of the audience for whom the programme is intended.* To demonstrate that care has been taken in this respect, the researcher should indicate the demographic and/or psychological bases on which the sample was selected. If rigorously controlled sampling is out of the question, the researcher must take care to *qualify the results accordingly.*

The sampling limitations of the Carter-Reagan study were stressed when its preliminary results were reported on CBC-Radio the morning after the debate. The external validity of its main findings was indicated eighteen months later by results obtained independently in the United States (Wingerson, 1982). Fortunately, the response task used in the Carter-Reagan study was a simple one, which the subjects were evidently able to fulfil with a high degree of reliability. Although the external validity of the study had been jeopardized by sampling restrictions, its internal validity was apparently high, each subject being considered as his or her own control in a series of multiple response comparisons.

Obviously, care must therefore be taken to ensure that the response task involved in a continual response study is within the intellectual and physical means of the subjects. As the following section shows, typical response tasks can often be too difficult for some viewers to handle.

COMPLEXITY OF THE CONTINUAL RESPONSE TASK

In common with other response analyzers, the PEAC system allows for the collection of continual responses on one of two bases:

1) nominal, categorical responses — such as CARTER, REAGAN, or DON'T KNOW, or
2) an interval, or quasi-interval scale of responses such as GOOD, FAIRLY GOOD, FAIRLY POOR, POOR.

It also provides two modes of push-button response: the *Reset* mode, in which the appropriate button must be depressed continuously in order for a response to be registered, and *the Latched* mode in which the current response is assumed to pertsist

Figure 1. *The Carter-Reagan Debate (Question: "Who, from one moment to the next, is winning the most votes?").*

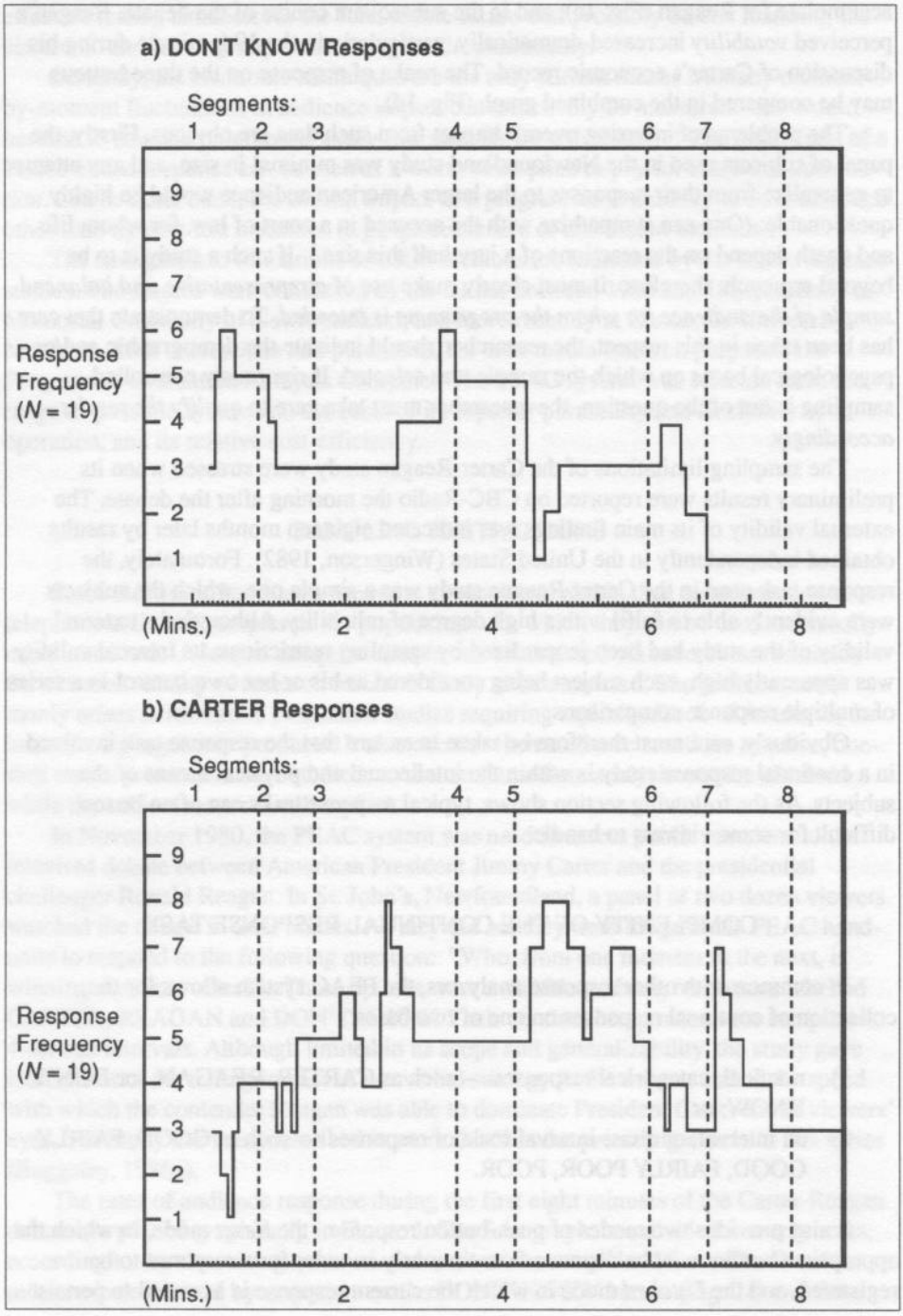
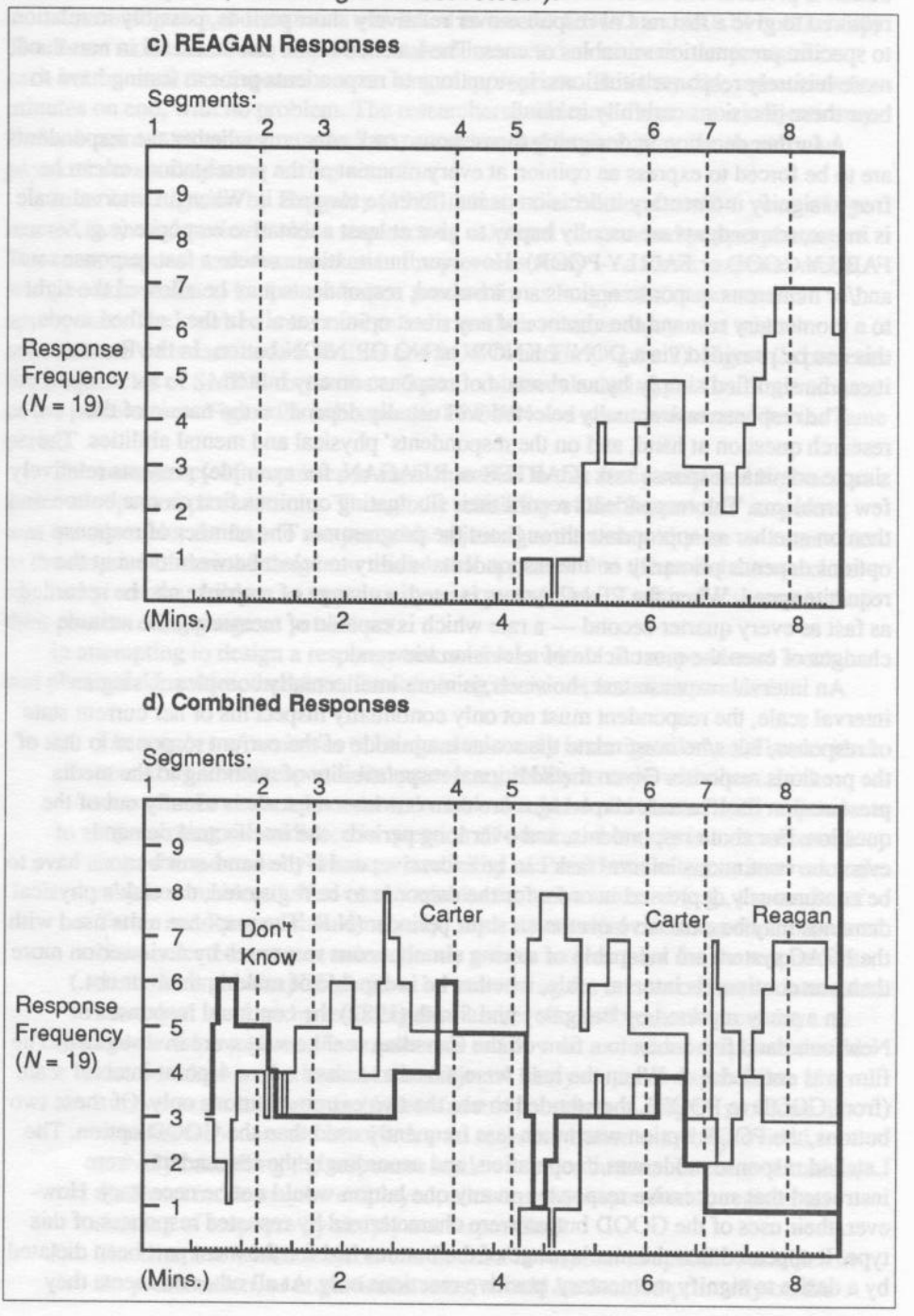


Figure 1, continued. *The Carter-Reagan Debate (Question: "Who, from one moment to the next, is winning the most votes?").*



— whether the appropriate button is currently being pressed or not — until a different button is pressed. The Reset mode can be useful in situations where subjects are required to give a fast rate of response over relatively short periods, possibly in relation to specific presentation variables or cues. The Latched mode can be useful in non-cued, more leisurely response situations. Instructions to respondents prior to testing have to bear these decisions carefully in mind.

A further decision in designing the response task concerns whether the respondents are to be forced to express an opinion at every moment of the presentation, or can be free to signify momentary indecision or indifference towards it. When an interval scale is in use, respondents are usually happy to give at least a tentative response (e.g., FAIRLY GOOD or FAIRLY POOR). However, in situations where a fast response rate and/or numerous response options are involved, respondents may be allowed the right to a momentary rest and the absence of any overt opinion at all. In the Latched mode, this can be provided via a DON'T KNOW or NO OPINION button. In the Reset mode, it can be signified simply by an absence of response on any button.

The response task actually selected will usually depend on the nature of the research question at hand, and on the respondents' physical and mental abilities. The simple nominal response task (CARTER or REAGAN, for example) presents relatively few problems. The respondents record their fluctuating opinions first on one button and then on another as appropriate throughout the programme. The number of response options depends primarily on the respondents' ability to select between them at the requisite speed. When the PEAC system is used. a change of response can be recorded as fast as every quarter-second — a rate which is capable of measuring the attitude changes of even the most fickle of television viewers!

An interval response task, however, is more intellectually complex. Using an interval scale, the respondent must not only continually inspect his or her current state of response, but s/he must relate the scalar magnitude of the current response to that of the previous response. Given the additional responsibility of attending to the media presentation itself, a task involving more than one interval scale is usually out of the question. For some respondents, and over long periods, the intellectual demands of even one continuous interval task can be excessive; and if the hand-unit buttons have to be continuously depressed in order for the response to be registered, the task's physical demands may be excessive over even short periods. (N.B. The response units used with the PEAC system are incapable of storing simultaneous responses by a viewer on more than one continuous interval scale, whether he is capable of making them or not.)

In a study reported by Baggaley and Smith (1982) the continual responses of Newfoundland fishermen to a film on the Canadian seal harvest were investigated. The film was entitled *A-I.* When the men were asked to assess it on a 4-point interval scale (from GOOD to POOR), they tended to use the two extreme buttons only. Of these two buttons, the POOR option was much less frequently used than the GOOD option. The Latched response mode was in operation, and accordingly the respondents were instructed that successive responses on any one button would not be necessary. However, their uses of the GOOD button were characterized by repeated responses of this type. It appeared that the men's usage of the buttons had for the most part been dictated by a desire to signify momentary, positive reactions only. At all other moments they

appeared to be either impartial about the film, or unwilling to register a negative reaction.

A second group of fishermen was asked to use the GOOD and POOR buttons only. The Reset mode was used, and the men were instructed to maintain the pressure on each button until the response was no longer appropriate. This they did, sometimes for minutes on end, with no problem. The researchers had learned that some subjects may find a 4-point interval response task too complex, and *to anticipate the mental and physical demands of each testing situation.*

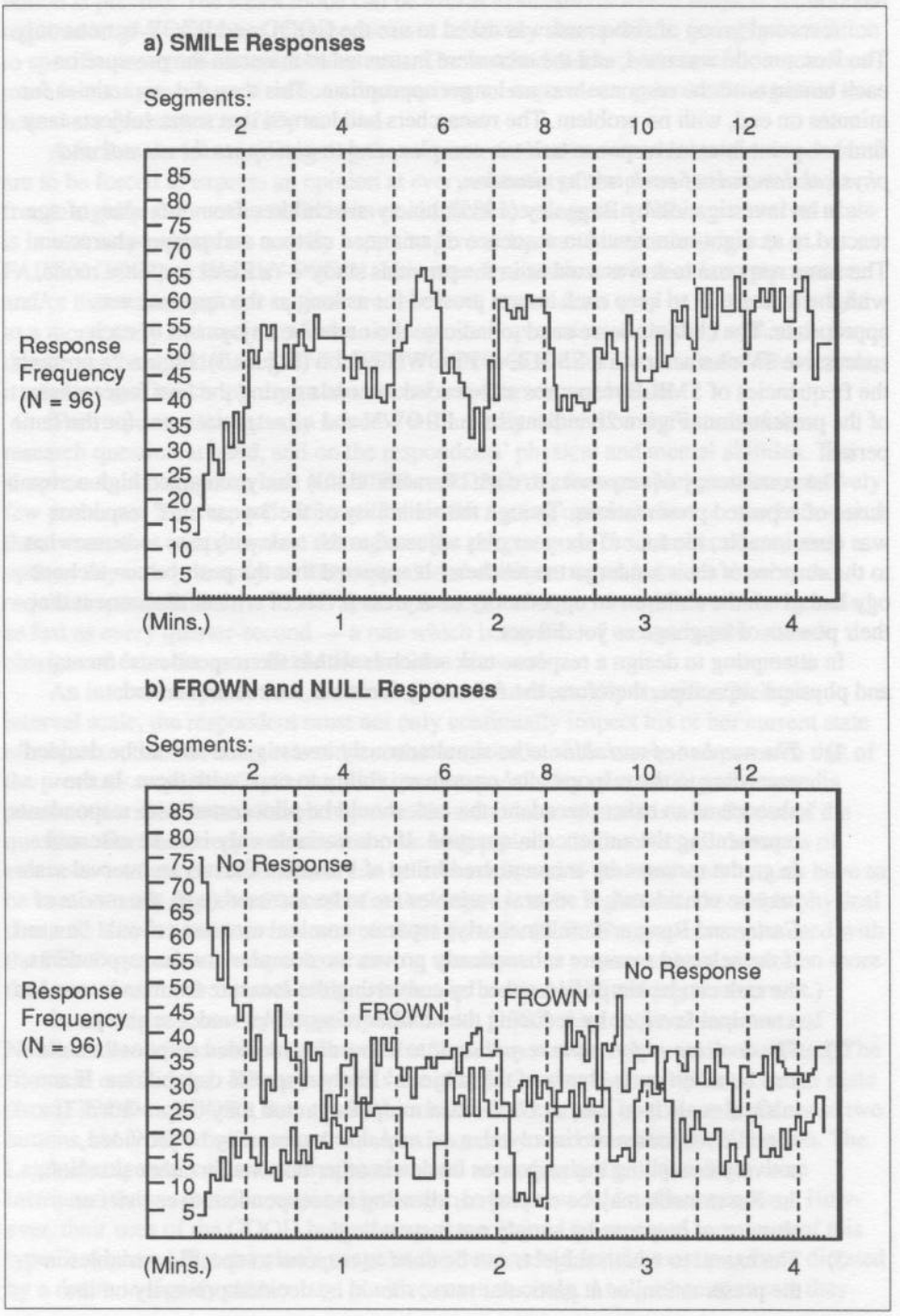
In an investigation by Baggaley (1985), ninety-six children from 3-6 years of age reacted to an eight-minute video sequence of animated cartoon and puppet characters. The same response task was used as in the previous study — a Reset response mode, with the instruction to keep each button pressed for as long as the response was appropriate. The children were cued to indicate their relative enjoyment of each successive TV character via a SMILE or FROWN button (Figure 2). Figure 2a presents the frequencies of SMILE responses at 2-second intervals during the first four minutes of the presentation. Figure 2b indicates the FROWN and *no response* rates for the same period.

The consistency of responses to each character in this study remained high across a series of repeated presentations. Though the reliability of the 3-year olds' responses was questionable, the four to six-year olds adjusted to the task with ease — somewhat to the surprise of their kindergarten teachers! It appeared that the push-button technology had given the children an opportunity to express levels of critical assessment that their powers of language as yet did not.

In attempting to design a response task which is within the respondents' mental and physical capacities, therefore, the following decisions are recommended.

1) *The number of variables* to be simultaneously investigated should be decided according to the respondents' customary ability to cope with them. In the absence of an exact precedent, the task should be pilot-tested with respondents representing the audience in question. If one variable only is to be assessed (e.g., the moment-by-moment credibility of President Carter) an interval scale may be considered. If several variables are to be assessed (e.g., the merits of Carter and Reagan simultaneously) separate nominal measures should be used. If the selected measure subsequently proves too complex for the respond&s, the task can be simplified either by converting the measure from an interval to a nominal level, or by reducing the number of variables under scrutiny.
2) *The decision to force rhe response* or to allow an undecided response should also be taken on the basis of the subjects' likely response capabilities. If an interval scale is in use, an *Undecided* midpoint button may be provided. If nominal responses are involved, a no *response* button may be provided, actively cancelling the responses made via other buttons. In either situation, the Reset mode may be employed, allowing the respondent to register an absence of response by simply not responding!
3) The extent to which subjects can be *cued* to *respond* to specific variables in the presentation, or at particular rates, should be decided primarily on the

Figure 2. *Preschool Children's Response to Cartoons and Puppets.*



basis of the presentation's length. At present, the tolerance of respondents for tasks of different lengths can only be judged intuitively. When the feasibility of a cued response task is in doubt, it should be avoided or the presentation shortened.

4) In all normal test situations, it is desirable for each respondent to be able to *refer to a visual display* of the most recent response on the hand-unit (as via the PEAC system). In this way, the respondent is reinforced in his or her efforts, and may also be reminded of the last response, as in Latched and interval situations.

SEQUENCE EFFECTS AND CONSTANT ERRORS

The data provided by continual response technologies can be both graphic and beguiling. Peaks and troughs in the moment-by-moment response profile invite instant interpretations of, for example, 'high visual interest', 'medium programme appeal', 'low presenter credibility' — and so on, depending on the response measure used. Such interpretations may be quite invalid. In one case, the writer had to restrain a TV producer from summarily firing the programme presenter in response to low rates of audience reaction that were observed during his appearances. It was pointed out that a low rating for visual appeal did not necessarily disqualify the presenter as a good educator. Conversely, a programme or programme presenter may receive a consistently high moment-by-moment rating, and yet be obviously failing in its attempt to fulfil the main programme objective.

Considered in isolation, the inferential value of continual response data is actually very low. In common with other forms of data gathered in sequence, they are subject to various types of psychometric error. When the continual ratings of a programme are generally positive, for instance, a momentary lapse in programme quality may not elicit the negative responses that it would otherwise: the segment will seem better in the sequential context than it would when judged on its own merits. When one programme segment follows others which are highly unpopular, on the other hand, its momentary ratings may suffer by association: it will seem worse than when judged on its own merits. These tendencies to over or under-estimate in a continual response task are identified as 'series' and 'time-order' effects (Woodworth & Schlosberg, 1961). The significance of sequence effects in PEAC system studies of reactions to advertising has been established empirically by Fenwick & Rice (1987): when advertisements were presented at the beginning of a test sequence, they were virtually always evaluated more positively than when screened later in the sequence.

The precise psychological meaning of CRM data is particularly difficult to interpret when the data are interval in nature. It is often unclear whether an interval response should more appropriately be interpreted as an absolute judgment on the scale, or as a relative one. A viewer's response on the FAIRLY GOOD button, for example, may be construed at its face value as representing an absolute judgment of 'fairly good'. But it might also represent a sudden, immense improvement in perceived quality to FAIRLY GOOD from VERY POOR, and a relative judgment whose correct

interpretation is identical to that of a shift from FAIRLY POOR to EXCELLENT. The viewer's current choice of buttons in a continual response task is dependent on relative as well as absolute judgmental forces in this manner.

Hand-units featuring a series of distinct buttons, as in the PEAC system, are actually less susceptible to psychometric error than other technologies demanding responses on an analogue dial. 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 a potentially infinite response scale do not ultimately yield more reliable measures of psychological impact, for they are subject to constant over- and undershooting errors known as habituation and anticipation bias respectively (Woodworth & Schlosberg, 1961). 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.

On all systems, however, the psychological meaning of momentary responses is obscured when responses are averaged across a group of respondents. Clearly, the attempt must be made to validate continual response data by referring *them to general criteria for programme effectiveness.* Examples of criterion-referencing strategies are given in the next section.

CRITERION-REFERENCING OF CONTINUAL RESPONSES

Criterion-referencing of continual response data can typically be achieved by:

1) comparisons between the responses of different viewing groups; (it may be critical, for example, that the responses given by women to a programme are more positive than those given by men); or
2) comparisons between moment-by-moment responses and a measure of overall programme impact as yielded by a pre- and/or posttest.

Criterion-referencing related to between-group comparisons may be appropriate in situations where a producer requires evidence of the programme elements which are capable of interesting one particular audience sub-group as opposed to others. For example, in the study by Baggaley (1985) of preschool children's responses to TV cartoon and puppet characters (see previous section), particular comparisons were made between the reactions of the boys and girls, and between those of English and French speaking children. When the continual responses of the boys were compared with those of the girls, sex differences in their preferences for particular characters emerged. No such difference was observed on the basis of the children's cultural background. The sponsors of the study, the National Film Board of Canada, received feedback about the types of TV character most likely to appeal simultaneously to both boys and girls.

In the study reported by Baggaley and Smith (1982), on the other hand, a

continuous measure of audience response was compared with measures of overall programme impact derived from pre- and posttests. The film in question concerned the Canadian seal hunt, a controversial object of protest by international conservationist movements; it aimed to teach seal fishermen ways of relining their sealing techniques and of increasing their financial yield from the hunt. The continuous measure of response was one of general approval towards the film, on a scale from GOOD to POOR. The overall measures related to shifts in attitude towards the seal hunt, and in learning about it, as measured from immediately before the film to immediately after it. If a positive continual response in such a situation were to be accompanied by minimal, or even undesirable overall effects on attitudes or learning, it would be obvious that the overall responses had greater validity as an educational index. The high continual responses would be either 'not high enough' in relation to the overall effect; or they could actually be quite irrelevant to it. Only when used as complementary to overall criteria, can continual response data have predictive meaning.

Particularly vital information in this study was gained from the responses of a group of Newfoundland high-school students. At first glance, their data seemed to indicate the type of disapproval shown towards it by the seal hunt protesters. However, on closer examination of their data, a totally contrasting interpretation was found to be tenable.

At a particular moment in the film, the killing of a seal was shown. At the same moment, a sudden shift was observed in the audience's continual responses towards the negative end of the approval scale. When a nominal response scale is used, response fluctuations of this type are apparent in terms of the number of audience members shifting from one button to another at a given moment, as in Figures 1 and 2 above. When a series of buttons representing an interval scale has been used, the levels of response may be assessed in terms of either: a) the frequencies of response on each button individually; orb) the average response on all of the buttons at once.

The shift towards disapproval by the high-school students on seeing the seal killed is apparent in Figure 3a. The graph shows a 50% drop in the number of students pressing the GOOD button at that particular moment in the film (i.e., during the eighth minute). The figure may be compared with Figure 3b, in which the group's average response on all four of the interval-scale buttons is presented — a more precise profile based on far more information. Both graphs are, of course, totally ambiguous with regard to the meaning of any particular moment-by-moment response. In this case, the researchers decided that the sudden response shift signified either distaste for the killing of the seals, or disapproval of the film for showing it, or a combination of both. In an attempt to determine which of these three interpretations was the most probable, the responses at this moment in the film were referred to the information about each respondent collected via the pre- and posttests.

Individual differences in responses to the killing sequence could thus be related,

1) to demographic information about the respondents (their age, sex, family background); as well as
2) to their prejudices about the seal hunt; and
3) to changes in their attitudes after seeing the film.

FIGURE 3. *Highschool Students' Responses to a Film about Sealing.*

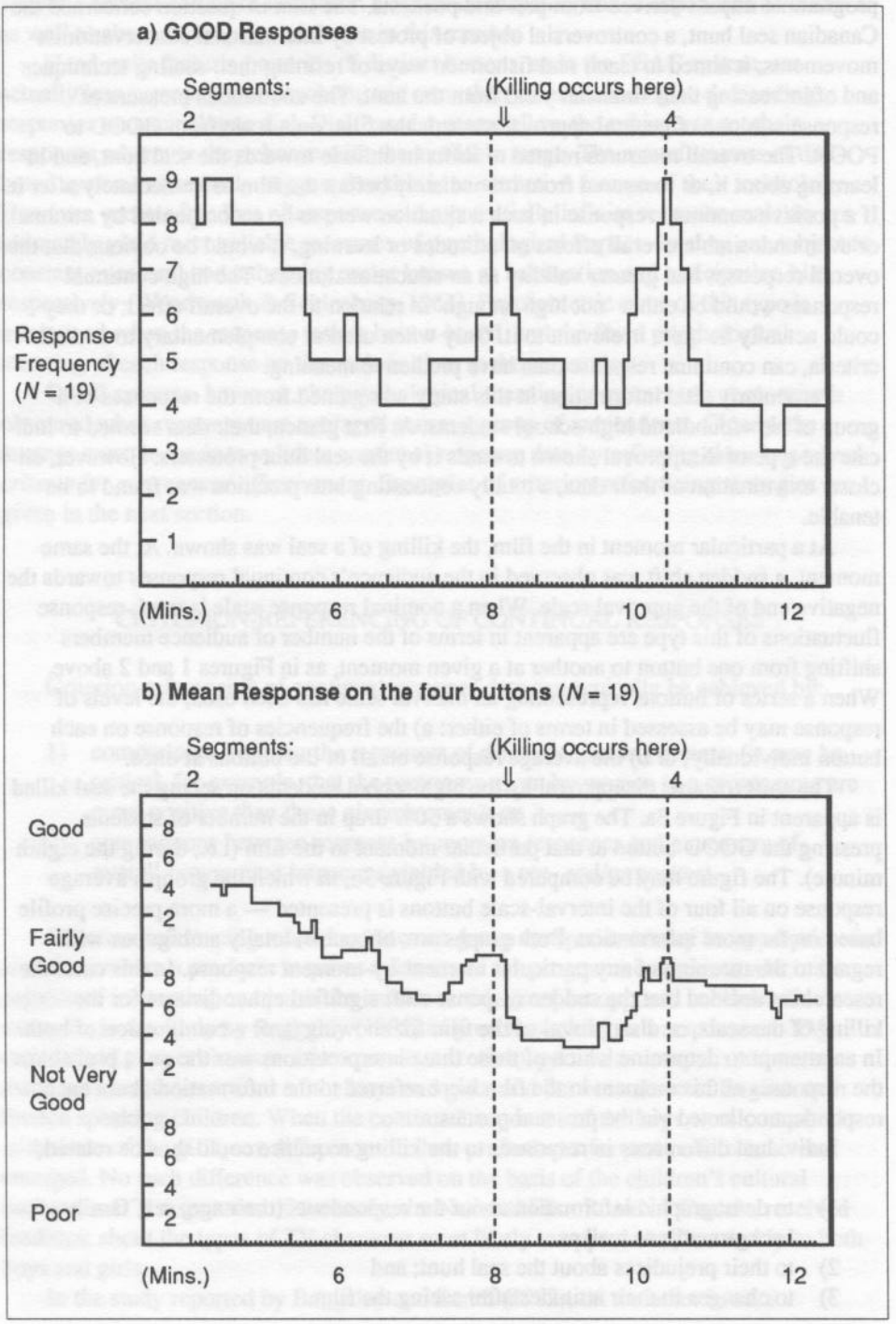
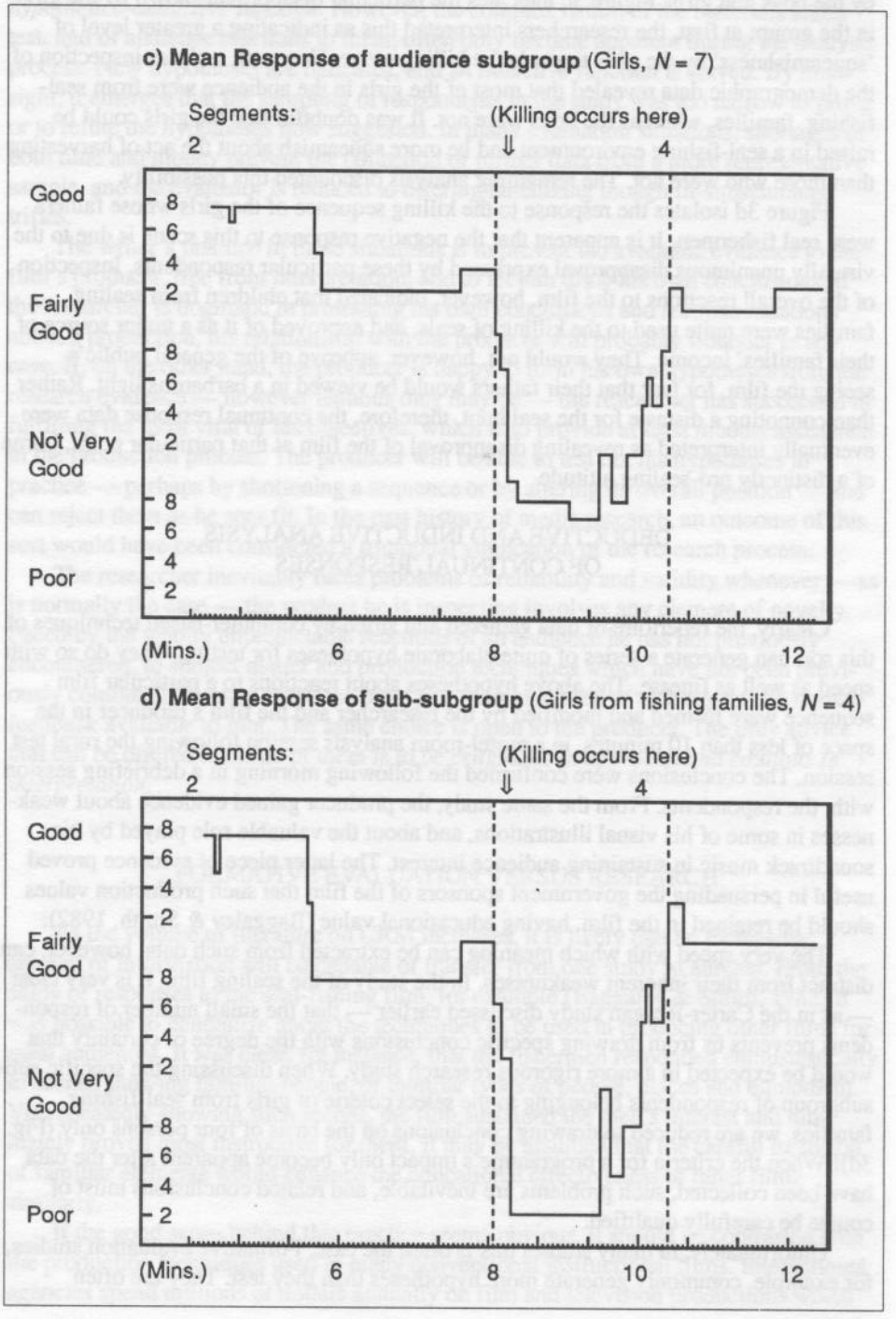


FIGURE 3, continued. *Highschool Students' Responses to a Film about Sealing.*



A marked difference was observed in the responses to the killing sequence given by the boys and girls. Figure 3c indicates the particular disapproval shown by the girls in the group; at first, the researchers interpreted this as indicating a greater level of 'squeamishness' by the girls towards the killing scene. However, a closer inspection of the demographic data revealed that most of the girls in the audience were from seal-fishing families, whereas the boys were not. It was doubtful that the girls could be raised in a seal-fishing environment and be more squeamish about the act of harvesting than those who were not. The remaining analysis discounted this possibility.

Figure 3d isolates the response to the killing sequence of the girls whose fathers were seal fishermen. It is apparent that the negative response to this scene is due to the virtually unanimous disapproval expressed by these particular respondents. Inspection of the overall reactions to the film, however, indicated that children from sealing families were quite used to the killing of seals, and approved of it as a major source of their families' income. They would not, however, approve of the general public's seeing the film, for fear that their fathers would be viewed in a barbarous light. Rather than connoting a distaste for the seal hunt, therefore, the continual response data were eventually interpreted as revealing disapproval of the film at that particular point, borne of a distinctly pro-sealing attitude.

DEDUCTIVE AND INDUCTIVE ANALYSIS OF CONTINUAL RESPONSES

Clearly, the repertoire of data gathered and sifted by computer-based techniques of this sort can generate a series of quite elaborate hypotheses for testing. They do so with speed as well as finesse. The above hypotheses about reactions to a particular film sequence were formed and modified by the researcher and the film's producer in the space of less than 10 minutes, in a motel-room analysis session following the rural test session. The conclusions were confirmed the following morning in a debriefing session with the respondents. From the same study, the producer gained evidence about weaknesses in some of his visual illustrations, and about the valuable role played by his soundtrack music in sustaining audience interest. The latter piece of evidence proved useful in persuading the government sponsors of the film that such production values should be retained in the film, having educational value (BaggaIey & Smith, 1982).

The very speed with which meaning can be extracted from such data, however, can distract from their inherent weaknesses. In the study of the sealing film, it is very clear — as in the Carter-Reagan study discussed earlier — that the small number of respondents prevents us from drawing specific conclusions with the degree of certainty that would be expected in a more rigorous research study. When discussing the specific sub-subgroup of respondents belonging to the select coterie of girls from seal-fishing families, we are reduced to drawing conclusions on the basis of four persons only (Fig. 3d)! When the criteria for a programme's impact only become apparent after the data have been collected, such problems are inevitable, and related conclusions must of course be carefully qualified.

Unfortunately, in many studies this is often the case. Formative evaluation studies, for example, commonly generate more hypotheses than they test. They are often

designed with specific hypotheses in mind, and may serve a hypothesis-testing, or *hypothetico-deductive* function. However, the complex nature of the materials under test, and of audience reactions to them, often only become apparent during the analysis process. New hypotheses are indicated, and *an inductive* function is served. By hindsight, it emerges that the sampling of respondents in the study was too narrow to prove or to refute the hypotheses now suggested. In many evaluation situations, shortages of both time and money prevent the collection of further data from a more representative sample, and the evaluator is reduced to offering tantalizing though ill-substantiated trifles.

The writer's practice in these situations is to present the available evidence to the film's producer, free from interpretation, and to let him draw his own conclusions. If the researcher is dogmatic in presenting his own conclusions and recommendations about a production, his relationship with the producer will probably flounder in any case. If, on the other hand, the producer is happy to form his own hypotheses from the research evidence — however tenuous they may be — the researcher has succeeded in fulfilling the most vital of his objectives, which is to provide at least modest assistance in the production process. The producer will be able to test out his hypotheses in practice — perhaps by shortening a sequence or by altering its overall position — and can reject them as he sees fit. In the past history of media research, an outcome of this sort would have been considered a triumphal vindication of the research process.

The researcher inevitably faces problems of reliability and validity whenever — as is normally the case — the product he is inspecting involves any element of novelty. Faced by the wholly unpredictable reactions of an audience he has not previously encountered, to subject matter and production techniques which have not been previously combined, he must make the most judicious sampling of techniques and human feedback available to him. The same choice is open to the producer. The only advice that can be offered to either of them is to be *pragmatic in approach and cautious in interpretation.*

FORMATIVE EVALUATION VERSUS RESEARCH

As the volume of literature on CRM increases, it is likely that an increasing amount of information will be capable of transfer from one study to another. From the study of responses to the seal-fishing film, for example (Baggaley Smith, 1982) it was possible to generalize about the techniques to be used in producing other films for rural audiences. It was clear, for instance, that the fishermen responded enthusiastically to scenes showing familiar people, or familiar types of people, places and equipment. Via repeatedly showing such scenes, the film sustained the men's interest and ultimately proved most instructive for them. It may be assumed that the careful inclusion of familiar elements would enhance the educational effectiveness of other films similarly.

If the good sense behind this practice seems obvious, it should be contrasted with the production techniques used in many conventional instructional films. International agencies spend millions of dollars annually on film and television productions which

are blithely assumed to meet the needs of their intended audiences. Lavish productions concerning health and work habits are released for a wide range of audiences, both educated and less educated. Rural audiences are constantly expected to identify with films centering around the unfamiliar activities and types of people found in urban communities; and the attention paid to pre- and pilot-testing of the films' educational impact is minimal.

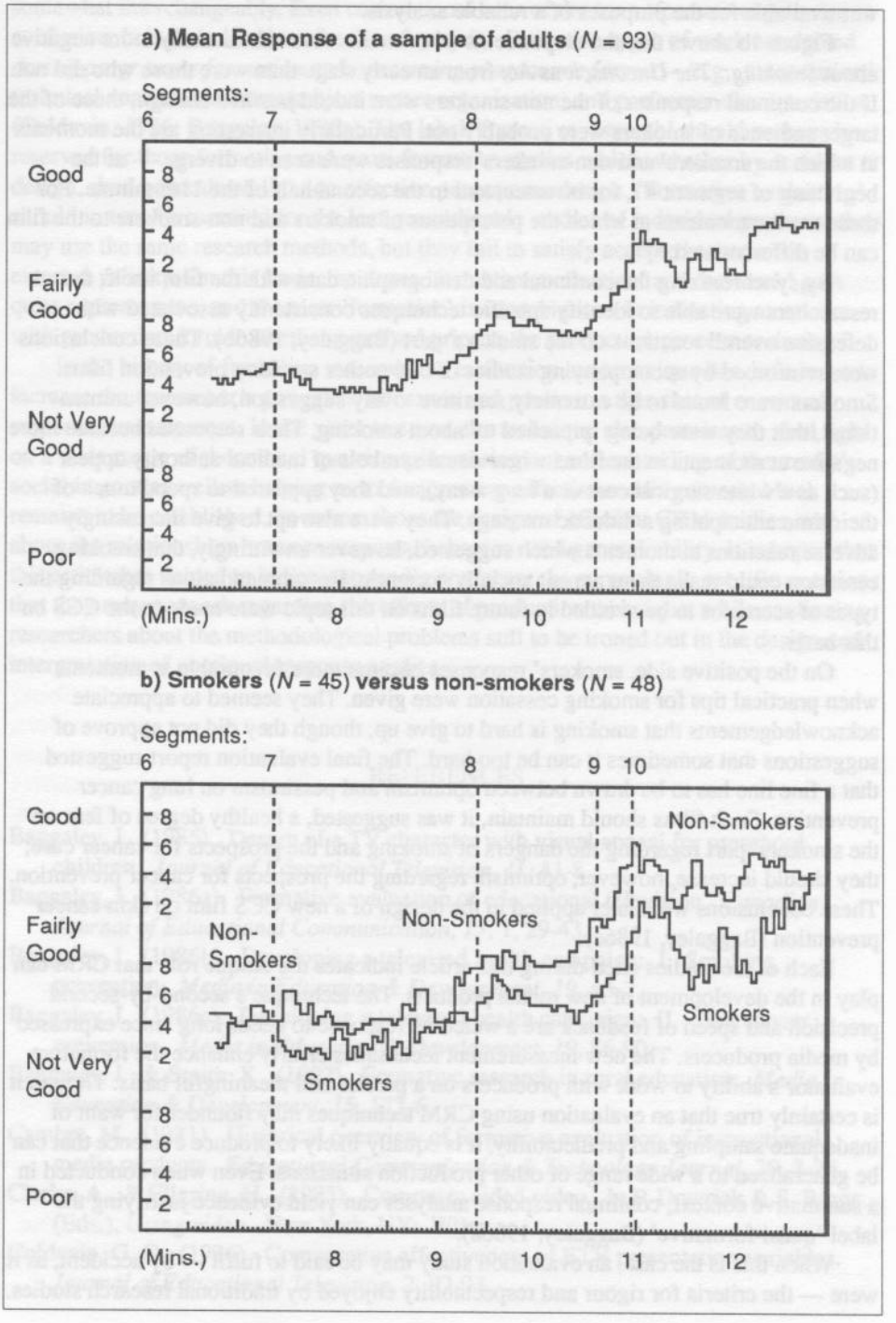
As the availability of funds for the production of educational media materials decreases, so the need for evaluation and improvement of their cost-effectiveness is intensified. In 1981, one national health agency was sufficiently concerned about the effectiveness of its educational media materials, that it embarked upon a detailed evaluation study of their impact upon a wide range of the intended audiences. The Canadian Cancer Society (CCS) had produced and distributed a wide range of cancer education films for all sectors of the Canadian public. It suspected that many of its educational materials — particularly those reliant on reading skills — were providing little or no benefit to the sectors of society with the most need for them. In areas such as lung and skin cancer prevention, the most needy sectors were perceived as the rural and 'functionally illiterate' communities.

The DEMO Project (Baggaley, 1986b, 1986c) was designed to investigate this possibility, and to recommend ways in which the impact of the CCS's public education programme might be improved. By evaluating the impact of specific films, the project aimed to derive generalizable research conclusions in this regard. In fact, the need for improvement of materials was found to be more severe than had been initially assumed. Male audiences generally were found to be defensive on cancer education matters. Their resistance to the types of film currently available to them transcended educational and social boundaries. The viewers receiving the least benefit from conventional smoking prevention films were those who smoke. Only the non-smokers were reinforced in the belief that smoking is unpleasant and should be avoided. After viewing some of the films — professionally produced by leading Canadian and American production houses — the smokers in the audience were more militant about their right to smoke than they had been beforehand.

Figure 4 reflects the responses of 93 people to one such film ***(Smoking: The Unconscious Act).*** In Figure 4a, a steadily increasing rate of approval is indicated by the average responses of the sample from one moment to the next on a 4-point scale. Particular segments of the film are seen as more or less effective on this basis. The usual problems of interpreting the graph are faced, of course. Although a high point in the film's perceived value is evident at the beginning of the 11th minute, there is no means of determining from the graph whether that moment is critical to the film's overall impact. The viewers' responses in general appear positive towards the film from the 11th minute onward, but one cannot tell from the graph alone whether they are *positive enough.*

Once the continual response data had been related to the independent demographic and attitudinal data, however, the meaning of the graph became gradually apparent. Breakdowns of the continual responses according to independent demographic data indicated few significant differences based on such factors as sex or age. The one variable which did affect the continual responses, however, was the audience's smoking

FIGURE 4. *Mean Responses to a Film on Smoking Prevention.*



behaviour; and in this case, fortunately, a statistically sufficient number of respondents was available for the purposes of a reliable analysis.

Figure 4b shows that the respondents who smoke were substantially more negative about Smoking: *The Unconscious Act* from an early stage than were those who did not. If the continual responses of the non-smokers were indeed *positive enough, those* of the target audience of smokers were probably not. Particularly interesting are the moments at which the smokers' and non-smokers' responses were seen to diverge — at the beginning of segment #7, for instance, and in the second half of the 1 lth minute. For these are the moments at which the perceptions of smokers and non-smokers to the film can be differentiated.

By synchronizing the continual and demographic data with the film itself, the researchers were able to identify specific techniques consistently associated with defensive overall reactions on the smokers' part (Baggaley, 1986b). These conclusions were reinforced by accompanying studies of three other smoking prevention films. Smokers were found to be extremely sensitive to any suggestion, however unintentional, that they were being 'preached to' about smoking. Their responses became more negative at moments in the films when visual symbols of medical authority appear (such as a white surgical coat or a lung x-ray), and they appeared to spend much of their time anticipating a didactic message. They were also apt to give increasingly adverse reactions at moments which suggested, however unwittingly, that smoking cessation could make them appear socially eccentric. Recommendations regarding the types of scenarios to be avoided in future films on this topic were made to the CCS on this basis.

On the positive side, smokers' responses became more favourable at moments when practical tips for smoking cessation were given. They seemed to appreciate acknowledgements that smoking is hard to give up, though they did not approve of suggestions that sometimes it can be too hard. The final evaluation report suggested that a fine line has to be drawn between optimism and pessimism on lung cancer prevention. Such films should maintain, it was suggested, a healthy degree of fear on the smokers' part regarding the dangers of smoking and the prospects for cancer cure; they should increase, however, optimism regarding the prospects for cancer prevention. These conclusions were later applied in the design of a new CCS film on skin cancer prevention (Baggaley, 1986c).

Each of the studies cited during this article indicates the unique role that CRM can play in the development of new media products. The technique's second-by-second precision and speed of feedback are a welcome response to needs long since expressed by media producers. The new measurement techniques clearly enhance the formative evaluator's ability to work with producers on a precise and meaningful basis. Though it is certainly true that an evaluation using CRM techniques may flounder for want of inadequate sampling and predictability, it is equally likely to produce evidence that can be generalized to a wide range of other production situations. Even when conducted in a summative context, continual response analyses can yield evidence justifying the label 'quasi-formative' (Baggaley, 1986a).

When this is the case, an evaluation study may be said to fulfil - by accident, as it were the criteria for rigour and respectability enjoyed by traditional research studies.

In previous literature, the terms 'formative evaluation and research' have been used somewhat interchangeably. Even conventional uses of the term 'formative' give rise to confusion, being used to describe the often quite distinctive types of work conducted *during programme formation* and *concerning programme format* — e.g., presentation/ technical variables; content/subject matter organization; and performer characteristics (Coldevin, 1976; BaggaIey, 1986a). The label 'formative research' should perhaps be reserved for those formative and quasi-formative studies which, whether by accident or design, shed generalized light on effective programme design. 'Formative evaluation' should in turn be reserved for the less generalizable studies of individual products; they may use the same research methods, but they fail to satisfy accepted criteria for external validity. On this basis, the terms 'formative evaluation' and 'research' gain quite separate uses, and the term 'formative' is unambiguous in indicating a concern with production format whether practised prior to the production process or during it.

In the history of formative research and evaluation, no technique has done more to increase the researcher's powers of inference and prediction than that of continual response measurement. In the years to come, the technique also promises to shed light on a wide range of theoretical questions, increasing our understanding of the media's social impact as well as helping us to design more effective media content. Much remains to be established, however, about the design of efficient CRM studies, and about the relationships between response behaviour and general ability. It is hoped that this article has helped to indicate to media producers the surprisingly specific questions that they may now ask regarding the effects of production technique, while cautioning researchers about the methodological problems still to be ironed out in the design and interpretation of continual response studies.

REFERENCES

BaggaIey, J. (1985). Design of a TV character with visual appeal for preschool children. ***Journal of Educational Television, 11,*** 41-8.

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

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

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

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

Cambre, M. (1981). Historical overview of formative evaluation of instructional media products. ***Educational Communication & Technology Journal, 29,*** *3-25*.

Clarke, A., & Ellgring, H. (1983). Computer-aided video. In P. Dowrick & S. Biggs (Eds.), ***Using video.*** New York, NY: Wiley.

Coldevin, G. 0. (1976). Comparative effectiveness of ETV presentation variables. ***Journal of Educational Television, 2,*** 87-93.

Corporation for Public Broadcasting (198 1). ***A comparison of three research methodologies for pilot testing new television programs.*** Washington, DC: Office of Communication Research, CPB.

Edel, R. (1986, Nov. 24). New technologies add dimensions to copy testing. ***Advertising Age,*** 520-3.

Fenwick, I., & Rice, M. (1987). ***Continuous measures of advertising response: Order effects.*** Unpublished manuscript, Faculty of Administrative Studies, York University, Toronto.

Lashley, K., & Watson, J. (1921). A psychological study of motion pictures in relation to venereal disease campaigns. ***Social Hygiene,*** 7, 181-219.

Malik, M. (1981). Biometric communication research for television. In J. Baggaley (Ed.), ***Experimental Research in TV Instruction*** (Vol. 4). Montreal, PQ: Concordia University.

Nickerson, R. (1979). Formative evaluation of instructional TV programming using the Program Evaluation Analysis Computer (PEAC). In J. Baggaley (Ed.), ***Experimental Research in TV instruction*** (Vol 2). St. John's, NF: Memorial University of Newfoundland.

Scriven, M. (1967). The methodology of evaluation. In B. Worthen, & J. Sanders (Eds.), ***Educational evaluation: Theory and practice.*** Worthington, OH: Charles A. Jones.

Woodworth, R., & Schlosberg, H. (1961). ***Experimental Psychology*** (3rd ed.), revised. London: Methuen.

Zirbes, L. (1924). Relation of visual aids to educational objectives. ***National Education Association Addresses & Proceedings*** (Vol. 62), 964-6.