DOCUMENT RESUME

ED 091 397 TM 003 587

AUTHOR Sanders, James R.; Cunningham, Donald J.
TITLE Techniques and Procedures for Formative Evaluation.
PUB DATE [74]
NOTE 54p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS Content Analysis; Data Collection; Design; Educational Assessment; Educational Needs; Educational Objectives; *Evaluation Methods; *Evaluation Techniques; *Formative Evaluation; *Literature Reviews; Program Design; Questionnaires; Sampling; *Summative Evaluation; Task Analysis; Test Construction
IDENTIFIERS *Product Development

ABSTRACT

After reviewing the literature, the authors defined a two dimensional framework comprising formative evaluation activity as one dimension and source of information as the other. Four types of formative evaluation activity were identified and defined. Three primary sources of information-internal, external, and contextual-were identified for consideration as the evaluator engages in the following four types of formative evaluation activity. The first section reviews a number of approaches to formative evaluation in the predevelopmental stage including sampling, Q-sort and task analysis. In the second section techniques for the formative evaluation of objectives are discussed, including questionnaires and surveys, delphi technique, and content analysis of documents. The third section deals with techniques for formative interim evaluation which may include collecting internal information such as descriptive information and processing critical appraisals, as well as describing physical specifications of the product. The fourth section deals with formative product evaluation in which a version of the complete product is produced. Rather than being discrete, this stage is continuous with evaluation of interim stages of the product. (Author/RC)

TECHNIQUES AND PROCEDURES FOR FORMATIVE EVALUATION

James R. Sanders[1] and Donald J. Cunningham

Indiana University

Sanders and Cunningham (1973) recently extended the writing of Scriven (1967) on the nature of formative evaluation applied particularly to the product development process. Formative evaluation was defined as the process of judging an entity, or its components, that could be revised in form, for the expressed purpose of providing feedback to persons directly involved in the formation of the entity. The authors defined a two dimensional framework comprising formative evaluation activity as one dimension and source of information as the other. Four types of formative evaluation activity were identified and defined as follows:

1. Pre-developmental Activities--formative evaluation work which occurs before formal product development has started. Formative evaluation tasks related to the evaluation of needs, tasks, other planning activities would fall into this category.

2. Evaluation of Objectives Activities--formative evaluation work directed at judging objectives in product development. The emphasis of work falling into this category would be on the provision of reliable information about the worth of goal statements produced by the product developer. Both logical and empirical evaluation strategies were proposed.

3. Formative Interim Evaluation Activities--formative work dealing with the appraisal of early product development efforts. Formal evaluation activities that would fall into this category were interim payoff evaluation work, interim intrinsic evaluation work and the evaluation of program or project operations. Informal evaluation activities, often unobtrusive, were also discussed.

4. Formative Product Evaluation Activities--formative evaluation work which focuses on the appraisal of a finished draft of the proposed product. Strategies such as validation studies, cost analyses, descriptive analyses and goal free evaluation directed toward a product draft would comprise this category.

Three primary sources of information were identified for consideration as the evaluator engages in the four types of formative evaluation activity listed above. The three sources were labeled and defined as follows:

1. Internal Information--information that could be generated by inspecting the entity itself. Included in this category would be descriptive information about and critical appraisals of the entity.

2. External Information--information concerning the effects of an entity on the behaviors of relevant groups. Student achievement after using a product or parental attitudes toward the objectives of a product would be information placed in this category.

3. Contextual Information--information concerning the conditions under which an entity is expected to function. Classroom environment, pupil characteristics and time of year are three examples of information that would fall into this category.

The two dimensions were crossed in a summary table that included evaluation techniques and procedures as cell entries. That table is reproduced here as Table 1.

Table 1

Summary of Techniques and Procedures
Appropriate for Formative Evaluation

SOURCE OF INFORMATION	FORMATIVE EVALUATION ACTIVITY			
	PRE-DEVELOPMENTAL	EVALUATION OF OBJECTIVES	INTERIM	PRODUCT
INTERNAL	logical analyses of needs: 1. cogency 2. consequences 3. higher order values empirical analyses: of needs: 1. group data: surveys scaling Q-technique semantic differential Delphi technique sentence completion 2. observation & expert opinion unobtrusive measures accreditation procs. category systems rating systems 3. analysis of documents unobtrusive measures content analysis	logical analyses: 1. cogency 2. consequences 3. higher order values empirical analyses: 1. group data: surveys scaling Q-technique semantic differential Delphi technique sentence completion 2. observation & expert opinion unobtrusive measures accreditation procs. category systems rating systems 3. analysis of documents unobtrusive measures content analysis	materials analysis guidelines content analysis analysis of learning structures group data (critical appraisal) expert opinion (including author) unobtrusive measures PERT PPBS system analysis	cost analyses materials analysis guidelines content analysis group data (critical appraisal) expert opinion unobtrusive measures
EXTERNAL		operationalization of objectives experimental try-out of goal statements	experimental and quasi-experimental design clinical methods quantitative naturalistic observation techniques unobtrusive measures	experimental and quasi-experimental design; hypothesis testing cost analyses GFE correlational analyses quantitative naturalistic observation techniques

Table 1
(cont)

	FORMATIVE EVALUATION ACTIVITY			
	PRE-DEVELOPMENTAL	EVALUATION OF OBJECTIVES	INTERIM	PRODUCT
CONTEXTUAL	needs assessment	context assessment (if no needs assessment results available)	literature reviews	unobtrusive measure group data perceived (on effectiveness of product) observation technique ATI procedures
			informal observation	context assessment (focus on external validity)

The purpose of this chapter is to expand on the earlier work by the authors by elaborating on selected techniques and procedures listed in the earlier work. Since space limitations do not allow the development of all techniques and procedures referenced in the previous work, the authors have selected those techniques and procedures which appear to be useful for formative evaluation. In addition, the authors have provided references to relevant techniques and procedures that were not selected for elaboration.

TECHNIQUES AND PROCEDURES FOR PRE-DEVELOPMENTAL FORMATIVE EVALUATION

Procedures and techniques for pre-developmental formative evaluation are often non-existent in typical evaluation systems, or, at best, they are very informal. Given the immediate need for production in most developmental projects, this situation is often explained away. But, it can never be reconciled when expensive errors are made during later stages of development. For this reason, we recommend the fullest amount of pre-developmental formative evaluation possible (within the constraints of scheduling, costs, and politics) using cheap approximations whenever formal, complete techniques and procedures are ruled unrealistic. The following are a few of the methods that the formative evaluator may want to draw on before development actually begins.

Reference was made in the earlier paper to needs assessment and needs (and object) evaluation procedures. It is

instructive in this regard to consider carefully procedures used by the National Assessment of Educational Progress (NAEP) project and the Institute for Social Research[2] (ISR) at the University of Michigan to identify refined techniques for accomplishing necessary pre-developmental activities. The techniques described here are ideals which can be quite expensive, but there is nothing to prevent the formative evaluator from adapting them to meet his needs. Two technical problem areas that have emerged frequently are those of developing a good sampling frame and of planning data analysis and reporting.

The NAEP sampling plans have been developed to meet two criteria: high accuracy in parameter estimation and low cost. The nation-wide probability sampling plan comprises a stratified multi-stage design. The parameter of interest is P_1, the proportion of the total number of persons in a certain subpopulation of the United States that answers an exercise in a certain way (e.g. P_1 = proportion answering 'yes'; P_2 = proportion answering 'no'; P_3 = proportion answer 'I don't know' to a three-option exercise). Each parameter is estimated by first estimating the total number of persons in the subpopulation, then estimating the number who would select each option on an exercise, and then expressing the estimate, P as a ration of the latter to the former. Such an estimate is called a 'combined ratio estimate' when applied to a stratified sample. The sampling plan includes listing units (small geographic areas, often counties, with a minimum size of 16,000 persons and easily

identified boundaries) as the major unit. Primary sampling units (PSU) are identified within each listing unit. In order to obtain a sample size of 2,000 responses per exercise (the n needed to obtain adequate precision), 208 PSU's with 10-12 observations on each exercise in each PSU used. The 208 PSU's are drawn from each of the four major regions of the United States (52 from each). PSU's are stratified on size of community, according to the census populations within each region, on income, and on geographic location within a region. The sample is selected by first drawing 1-2 PSU's per stratum randomly without replacement. Then, for the in-school sample, students from each PSU are listed. Sometimes large schools could contain students from several PSU's. Since a constraint of using at least two schools per PSU is placed on the sampling procedures, it has been important to associate students within the PSU with their schools. Approximately 250-350 students are expected to be within each PSU for each age group. Those students who actually participate in National Assessment are randomly drawn from the 250-350 students in each age group. For the out-of-school sample, PSU's are subdivided into secondary sampling units (SSU) which are clusters of 35-40 housing units. This procedure is very similar to the ISR procedure described below. Ten SSU's are randomly drawn from each PSU and are expected to yield 12.5 adult respondents on the average.

The Institute for Social Research also uses a multi-stage sampling plan for most of its large studies. The steps of the sampling procedure progress through various stages of selection

going from larger to smaller areas. These steps parallel closely those used by NAEP. Briefly, the ISR sampling plan leads to the identification of primary sampling units (PSU), usually counties or metropolitan areas, first. The PSU's are then stratified by relevant dimensions, such as urban versus rural areas, income of areas, and so on. A total of 74 PSU's are then randomly drawn from the strata, proportionate to the total number of PSU's in each stratum. The 74 selected PSU's are then subdivided into smaller areas called sample places and each sample place is subdivided into chunks which are areas within a sample place which have identifiable boundaries (e.g. township, city block, an area bounded by identifiable roads, streams, etc.). Several chunks are then selected randomly from each sample place for the sample and dwelling units are then identified within the selected chunks. It is here that the ISR procedures begin to differ from those of NAEP, the reason being the different purposes of data collection for the two projects. A final step in the ISR sampling procedure is the random selection of 3-4 dwelling units within a chunk, called a segment, and these dwelling units are used in the study.

The value of using multi-stage scientific sampling procedures for collecting survey data should not be underestimated. Obtaining precise estimates of relevant human parameters is an essential part of quality (precise?) product development. While the above discussion has not been prescriptive by any means (considering the almost infinite number of variations of basic multi-stage sampling plans,

context-free recommendations are virtually impossible), we would argue for the adaptation of sampling plans already in use for pre-developmental formative evaluation.

Reporting procedures for survey data used by NAEP provide a useful model for reporting pre-developmental formative evaluation activities. It is most informative to the reader to provide the complete item along with estimates of the proportion of persons in each subpopulation who would choose each item. It is also important to provide normative data for the subpopulation which can be used to interpret the reported parameter estimates. For example, data reported by geographical region on an overlap item given to respondents aged 13, 17 and adult, should include response estimates for the entire population (over all geographical regions) as well as data by age reported side by side for comparison purposes. It is recommended that all planning data (on needs, objectives, etc.) be reported alongside data on relevant referent groups (norms). Descriptive statistics and parameter estimates (along with standard errors) are the most useful data reduction procedures to use at this stage of the development of large-scale assessment procedures.

One technique for evaluating needs and objectives frequently mentioned in many recent papers on formative evaluation is the Q-sort. Methods for collecting appraisal or judgmental data from relevant groups of persons on simply and tersely stated needs or objectives is essential in pre-developmental and

objectives formative evaluation. Fortunately, the procedures developed by William Stephenson and labeled Q methodology are most appropriate. The Q technique is the logical operationalization of Stephenson's theoretical Q methodology. Briefly, a list of need statements or goal (objective) statements may be assigned numerals, placed on cards and given to persons to rank order according to some predetermined rules. The ordinal data that result from the sorts may then be analyzed to yield a number of useful statistics such as:

1. Consistency or homogeneity of ranking within a group of persons (answering the question of how much do people agree on their perceptions of the needs or objectives).
2. Overall (and subgrouped) rankings (or sets of priorities) on the list of needs or objectives (and also the variance for each need or objective statement).
3. Differences in ranking profiles among groups of persons (e.g. a summary of differences among a school board, the school teachers, the school administrators and parents on the priorities or values assigned to a list of needs or objectives).
4. Clusters of needs or objectives as ranked by a given group of persons.
5. Clusters of persons as they rank needs or objectives (e.g. Do Republicans versus Democrats cluster respectively on their priorities?).
6. Similarity of the distribution of rankings by a group of persons to an ideal or criterion distribution.

There are two basic types of Q-sort, each with a particular use: structured and unstructured. Structured Q-sorts are those that include a set of rules whereby a certain number of cards (needs or objectives) must be placed in each of a certain number of piles (e.g. left-hand piles for most valuable and right-hand piles for least valuable needs or objectives). Here we are forcing the sort into a predetermined distribution, according to some theory. Unstructured Q-sorts are those used where there is no underlying theory and we ask a person to merely place the cards into a predetermined number of piles according to his own perceptions of where they should be placed. In essence, we are saying here, "Let the cards fall where they may."

The procedures used to collect Q-sort data generally follow these steps:

1. Place unambiguous needs or objectives statements on cards, one to a card. Theoretically, at least 75 but no more than 140 items should be sorted.
2. Shuffle or randomly order the cards and give them to a person to sort. The same random order should be given to each person.
3. Sort the cards into some predetermined distribution. Usually 7-13 piles of cards are used, but this can be modified, depending on the needs of the investigator. For example, if 80 items were to be sorted into a quasi-normal distribution, the following rules might be set: Sort the cards into 9 piles with the number in each pile

set as follows:

4 6 10 12 16 12 10 6 4

The left-most pile represents most valuable needs or objectives and the right-most pile represents least valuable needs or objectives.

4. Collect the cards as sorted by the person and assign ranks to the cards in each pile (e.g. one to cards in left-most pile and 10 to cards in right-most pile).

5. Calculate desired statistics on resultant data.[3]

This technique has been used by numerous formative evaluators in Regional Labs and R & D centers, schools and universities for collecting judgmental data necessary in planning for product development.

A set of procedures which were not listed in the earlier paper by Sanders and Cunningham, but which should be essential to systematic product development are those that fall under the rubric of task analysis. This activity has been described in excellent discussions by Davies (1972) and Thiagarajan, Semmel and Semmel (1973). Early work by developers of programmed instruction has also contributed greatly to the refinement of task analysis techniques. This activity is not clearly evaluative since the judging process is not involved, but task analysis (like objectives writing) is often a function assigned to the evaluator by his clients.

There is some uncertainty about whether this activity is appropriate at the pre-developmental stage of product development

(we argue later that it is definitely an appropriate interim formative evaluation activity). We have included a discussion of task analysis here because we feel that at a point when global "needs" and "goals" are the only existing descriptions of the final product, it is worth partitioning these global outcomes into component parts. This activity could then lead directly into the preparation of interim and terminal objectives which have cogent bases for their existence. Because of the uncertainty associated with the appropriateness of task analysis activities at this point, we have discussed the technique both here and in the formative interim evaluation section of this chapter paper. Task analysis techniques are involved primarily with the prescription of the prerequisites and conditions under which behaviors may be developed and a description of the behaviors which comprise a given performance. Thiagarajan, Semmel and Semmel list the following steps for performing a task analysis:

1. Specify the main task [or performance]. This statement should indicate what the subject is to do upon the use of a given product and the situation in which he is to perform.
2. Identify subtasks. These statements should include the skills that the subject must possess in order to demonstrate the criterion performance.
3. For each subtask, identify sub-subtasks which contribute to that subtask.
4. Terminate reduction of tasks into subtasks when the subtasks are equivalent to the subject's entry behavior.

An example of such an analysis is found in Figure 1.[4]

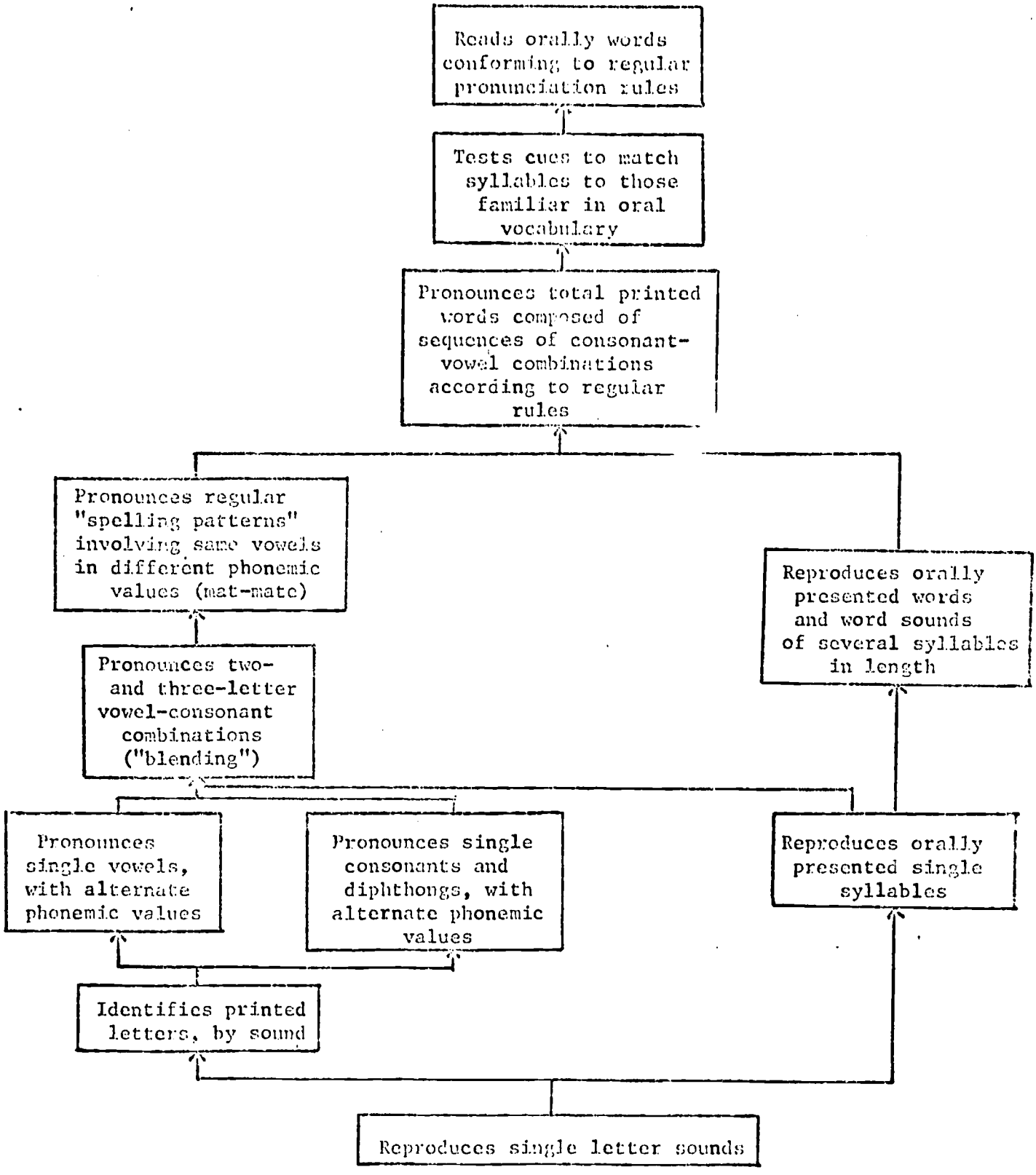

FIGURE 1. A learning hierarchy for a basic reading skill ("decoding").

Davies suggested that task analysis approaches such as the one described above are but one of six different approaches. The six categories of task analysis are:

1. Task analysis based upon objectives. This method includes the specification of instructional objectives and the specification, for each objective, of the type of behavior (e.g. knowledge, comprehension, receiving, responding, etc.) required for each.
2. Task analysis based upon behavioral analysis (above).
3. Task analysis based on information processing. This method includes a prescription of information to be processed for the performance to be mastered. Considerations of cues, manipulations to be made, feedback, etc. are central to this method.
4. Task analysis based on a decision paradigm Underlying decisions which must be made to perform a given task are analyzed and decision chains and procedures are provided.
5. Task analysis based on content structure. This method includes the identification of rules and examples invalued in the task, the presentation of these rules and examples and the discussion of relationships between them.
6. Task analysis based on vocational schemes. This method involves the reduction of a performance into jobs, duties, tasks and task elements.

Davies noted that these six categories are not mutually exclusive, but do suggest central elements of different approaches to the

problem of task analysis. The state of the art in this area is such that a considerable amount of research is still needed to expose the utility of each of task analysis approach for use in quality product development. While the approaches are quite promising, the methods of task analysis are still evolving. They should be refined by formative evaluators and data should be presented on the relative payoff of each.

While task analysis and learner analysis are not clearly evaluative functions (for the same reasons that needs assessment and objectives generation activities are not evaluative), they are evaluation-related functions that are essential parts of the development process. Often formative evaluators are called upon to perform such functions, and, as such, they should be techniques and procedures which the formative evaluator has in his repertoire. The techniques and procedures used to evaluate needs or objectives (as outlined by the authors in the earlier paper) are appropriate for the evaluation of task analysis and learner analysis results.

ADDITIONAL REFERENCES:[5] Surveys (Herriott, 1969; Institute for Social Research, 1969; Oppenheim, 1965), Scaling (Torgerson, 1958), Q-technique (Stephenson, 1953), Semantic differential (Osgood, Suci, and Tannenbaum, 1957), delphi technique (Helmer, 1967), accreditation procedures (NSSE. 1969, 1970), observation techniques and category systems (Simon and Boyer, 1970; Wolcott, 1968; Burnett, 1968, 1969), rating systems (Lawson, 1973),[6] content analysis (Berelson, 1952; Guttentag, 1971).

TECHNIQUES AND PROCEDURES FOR THE FORMATIVE EVALUATION OF OBJECTIVES

The importance of empirical methods in evaluating objectives was noted in the earlier paper by Sanders and Cunningham. Stake (1970) suggested three categories of judgmental data that might be collected in evaluating objectives: group data, expert opinion and observation data, and document analysis data. Two essential strategies for collecting group data include the use of survey questionnaires and the delphi technique. Student opinions and content analyses provide valuable information from "experts" and documents respectively.

The development of questionnaires is the most critical and possibly the most underemphasized part of survey inquiry. A common attitude among many evaluators and researchers is the predisposition to write quickly a list of questions to be answered, put them on a form and call the resulting instrument a developed questionnaire. In reality, if the questionnaire has not undergone critical appraisal before being sent to potential respondents, little usable information will be yielded. The early stages of survey design comprise decision-making about the aims of the study and identification of hypotheses to be tested or questions to be answered. Talking to experts and reviewing literature related to the evaluation focus should enable the evaluator to get a feel for the problem. After deciding on the questions to be answered, it is important to consider the analyses, results, etc. needed to answer the

questions. At that point, the evaluator should be able to infer the questionnaire questions that are to be asked and how they should be quantified. Criteria that could be used to evaluate draft versions of the questionnaire include:

I. Question Sequence
   - A. Are later responses biased by early questions?
   - B. Is the questionnaire attractive and interesting? Does it start off with easy, impersonal questions?
   - C. Are leading questions asked? Is there a logical, efficient sequencing of questions (e.g. from general to specific questions; use of filter questions when appropriate)?
   - D. Are open/closed ended questions appropriate? If closed, are the categories exhaustive, mutually exclusive? (Could ordinal or nominal data be collected as interval data?)
   - E. Are the major issues covered thoroughly while minor issues are passed over quickly?
   - F. Are questions with similar content grouped logically?

II. Question Wording
   - A. Are questions stated precisely? (Who, what, when, where, why, how?)
   - B. Does the questionnaire assume too much knowledge on the part of the respondent?
   - C. Are double questions asked?
   - D. Is the respondent in a position to answer the question, or must he make guesses?
   - E. Are definitions clear?
   - F. Are emotionally tinged words used?
   - G. Are technical terms, jargon, slang, words with double meanings avoided?
   - H. Are the methods for responding consistent?
   - I. Are the questions impersonal?
   - J. Are the questions short?

III. Establishing and Keeping Rapport
   - A. Is the questionnaire easy to answer?
   - B. Is little respondent time involved?
   - C. Does the questionnaire look attractive? (e.g. lay-out, quality of paper, etc.)
   - D. Is there a 'respondent orientation?'
   - E. Is the questionnaire introduced with an explanation of purpose, sponsorship, method of respondent selection, anonymity?

IV. Instructions
- A. Is the respondent clearly told how to record his responses?
- B. Are instructions for return due date and procedures included?

V. Technical Quality
- A. Validity
  1. Are second information sources used as cross-checks (interviewer ratings, other findings, etc.)?
  2. Are responses of like respondents (e.g. husband/wife) checked?
  3. Have content experts read pilot versions of the questionnaires?
- B. Reliability
  1. Are factual questions reasked?
  2. Are phoney items used?
  3. Are respondents reinterviewed?
  4. Have responses been checked for logical consistency?
- C. External Validity
  1. Are non-response bias checks planned?

An excellent annotated bibliography on the design, construction and use of questionnaires for inquiry is provided by Potter, et al. (1972).

A variant of survey procedures for collecting judgmental data is the delphi technique. This technique makes use of a panel of experts who are mailed a set of questions to which they respond independently. A follow-up questionnaire reports a summary of the original responses using the median and inter-quartile range as descriptive statistics for the responses to each original question. Each panel member is then asked to reconsider his first responses and revise them if he so desires. If his second response is outside the interquartile range, he is asked to justify his deviation from the majority judgment.

In the third round, the second round responses are summarized and a summary of the reasons provided for deviant positions is also included. Each panel member is asked to reconsider his second round responses given the results and reasons yielded from that round. A respondent who desires to remain outside the interquartile range on the third round is asked to present his reasons. This iterative procedure can continue for several more rounds after the third, but the payoff begins to diminish quickly. On the final round, panel members are asked to revise their responses one last time given the results and arguments yielded by the previous round. This procedure has been used in management to attain consensus judgments from a panel of experts. Often the results have been less than spectacular due to weaknesses inherent in the process, but on many occasions useful results have been obtained. This is a procedure that the formative evaluator may find useful in the early stages of product development when committments on selected developmental goals must be made.

Abedor (1971) suggested procedures for collecting judgmental data about objectives from representatives of the target population for a product. His procedures could be adapted so that subjects are given an objective or a list of objectives and are asked to react to them as behaviors that the subjects could be asked to demonstrate after using the product. A valuable lesson here is that members of the target population are 'experts' who are often overlooked in formative evaluation.

In addition, they are one of the most critical and insightful audiences available to the evaluator. Since they will be suffering the consequences of bad development in the long run, they have something to lose by not providing feedback to the evaluator.

For the analysis of documents for collecting judgmental data about objectives, content analysis procedures have much to offer. Content analysis aims primarily at the objective quantification of content classified using a system of categories and explicitly formulated rules. The categories should be developed to fit the questions to be answered by the data and they should be mutually exclusive and exhaustive. Coding units (e.g. words, themes, paragraphs, etc.) are what the content analyst actually counts and places within the categories. A sample set of categories into which themes contained in newspaper articles dealing with sex education could be tabulated might be:[7]

Newspaper: Date: Story Source:

-		+	
Expressions of opposition to sex education		Expressions favoring sex education	
Actions in opposition to sex education		Actions in support of sex education	
Statements supporting opponents of sex education		Statements attaching opponents of sex education	
Statements attaching proponents of sex education		Statements supporting proponents of sex education	
Statements listing opponents of sex education		Statements listing proponents of sex education	
Provisions of alternate plans		Statements opposing alternate plans	
Some other plan satisfactory		Authorities insist on current objectives	
Miscellaneous-		Miscellaneous+	

0		Other themes	
School Board to Discuss Issue			
School Board Vote to be close			
Possible Areas of Compromise			
Miscellaneous			

Content totals	Headline	Headline Content
+ _____	Head Size _____	(+1, -1, or 0)
- _____	Location on Page _____	
0 _____	Length _____	
	Total Score and Direction _____	

The uses of this technique for collecting judgmental data on objectives are many. Thematic analyses of board meetings or editorials in professional journals or word counts on federal policy statements can identify and clarify value data that are unavailable from any other source.

TECHNIQUES AND PROCEDURES FOR FORMATIVE INTERIM EVALUATION

At this point in the product development process, "pieces" of the intended final product are beginning to emerge. A film maker, for instance, often begins by constructing a series of verbal descriptions of the visual stimuli which he intends to film, coordinating that description with a preliminary version of the stimuli to be presented on the sound track (if any). A frequent next step is the construction of a "story board," or simulation of the visual and oral stimuli with hand drawn pictures or photographs serving for the visual stimuli. Some

film producers use the relatively less expensive videotape medium to film initial versions of their film for "debugging" purposes. Film is a particularly difficult medium to revise once it has reached the finished product stage in that changes are likely to cost as much as the product itself. So it is extremely important to locate points in the interim stages in development of filmed materials where evaluative information can be provided concerning potentially useful revisions.

But, the fact that revisions of a finished product are less costly in some other media should not obscure the usefulness of seeking evaluation at the interim stage of product development. Most textbook authors begin by constructing some sort of topical outline, chapter summary, etc. Rough drafts of chapters often undergo several revisions based upon feedback from colleagues, students, spouses, secretaries and anyone else who the author can coerce into reading his smudged drafts. Small scale tryouts of each chapter as it rolls off the pencil are often undertaken. The point is, of course, that in the development of nearly any product many opportunities exist prior to the completion of the initially satisfying version of the complete product for evaluative information to be collected. The particular techniques useful for formative interim evaluation of various media will differ somewhat from medium to medium but many general principles can be noted.

Formative interim evaluation information can involve collecting internal information such as descriptive information

and processing critical appraisals. Descriptive information refers to the objective information which can be generated by inspecting the pieces or preliminary versions of the product. Critical appraisals are judgments made concerning the pieces by representatives of concerned populations (e.g. experts, parents, students, etc.). Each of these will be discussed in turn.

The intent of collecting descriptive information is to describe fully and completely what is, not what should be. A comprehensive characterization of what is will aid greatly in making judgments and in determining where to revise once some deficit is identified.

One type of descriptive information is physical specifications which is simply a description of the primary "tangible" characteristics of the product consisting in large part of media characteristics. This type of information is best collected by means of a checklist which includes the majority of the characteristics upon which products can vary. These characteristics are usually media specific in that any general purpose checklist would be impossible to construct. Some sample characteristics are listed in Table 2 using programmed instruction as an illustrative medium.

Table 2

Sample Items from a Checklist for Evaluating
Descriptive Characteristics of a Programmed Textbook

1. Pre-test provided? _____ Yes _____ No

2. Objectives listed? _____ Yes _____ No

3. Confirmation procedure. Check one.

_____ Knowledge of results provided on same page, students asked to shield answer.

_____ Knowledge of results on another page of text.

_____ Knowledge of results provided in separate booklet.

_____ Knowledge of results not provided.

_____ Other (Please specify).

4. Response requirement (intended):

_____ Overt constructed.

_____ Covert constructed.

_____ Overt selection.

_____ Covert selection.

_____ Other (Please specify).

5. Can student alter response requirement? _____ Yes _____ No

6. Blackout Ratio.

_____% of material could be blacked out.

Some points in the sample checklist note whether certain features are present or absent in this product. (See points one and two.) Other points identify the type of procedure that is employed for features which are present invariably (or nearly so) for every product of this type. That is, nearly every program requires some type of response from the student and nearly every one provides knowledge of results in some form. But different programs require different types of responses and use different methods of providing knowledge of results. (See points three and four.)

Similar checklists could be developed for any medium or combination of media. Checklists have, of course, been developed and used for many years (e.g. Edmonson et al., 1931; Hoban, 1942) but these checklists require rather global judgments by the user and are likely of more use to the summative evaluator. One would expect that a number of generally accepted checklists of potentially useful descriptive information for formative evaluations would be available for instructural products of many types but, to our knowledge, this is not the case. Each developer, if he concerns himself at all with descriptive information, rediscovers the wheel, so to speak. The disadvantage of such a state of affairs is that the developer may not be aware of potentially useful types of descriptive information.

↕ double space

One method which offers promise for describing product content is content analysis. The content analysis procedures discussed earlier in this ~~paper~~ chapter are also appropriate for formative

interim evaluation. Berelson (1954) defined the technique as a "research technique for the objective systematic and quantitative descriptions of the manifest content of communication." In addition to the content analysis techniques described earlier, an introduction to content analysis may be found in Kerlinger (1973) while more advanced treatments of the topic may be found in Budd et al. (1967) and Halste (1969). Grobman (1972) has provided a useful discussion of the uses of content analysis in formative and summative evaluation although her discussion seems more oriented toward summative evaluation.

Content analysis, however, does not lend itself easily to a consideration of the relationship among concepts in the subject matter. The learning structure analysis of Gagne (1970) is very useful in this regard. This technique, also described earlier in this ~~paper~~ chapter under task analysis procedures, is relevant for the formative evaluation of interim products. Gagne (1970) presents many examples of learning hierarchies and the technique seems to offer many advantages.

The construction of learning hierarchies is quite time consuming, however, and the construction of a learning structure is no guarantee that it is "correct." In essence, the learning structure is a logical analysis of the objective and, as is well known, logic does not always simulate reality. Skills which are presumed to be subordinate to a particular objective may turn out not to be or the sequence of subconcepts may prove to be wrong. Learning hierarchies are, in essence, hypotheses

concerning the content, hypotheses which can only be confirmed empirically. The usefulness of a particular learning hierarchy will depend upon how well it fits the reality of the situation. Many subject matters do not lend themselves to hierarchical analysis. In other words, this, and systems similar to it, do not possess unlimited applicability but they should prove useful in many situations.

The reader may, at this point, be wondering whether the collection of descriptive information is really necessary. Many would argue that all that really counts is how well the product works, not what it consists of. The trouble with that argument is that not all products work, especially in first draft form. When a product fails to perform as expected, explanations must be found. An adequate inventory of descriptive information will assist greatly in locating the points at which the product needs revisions. The particular information collected will depend on many complex factors: cost, utility, past experience, etc. As such information is collected more often, the collection should become easier. Instruments such as a checklist will already be constructed, past content analysis systems already "debugged," etc. The reader is also referred to the CMAS, Eash, and Tyler and Klein analysis procedures referenced in the earlier paper by Sanders and Cunningham (1973) as well as procedures provided by the Educational Product Information Exchange (EPIE) (1972).

↕ double space

The activity of critical appraisal also refers to inspecting the product itself. Critical appraisal is not the appraisal of the effects of a product upon people who are using the product but upon people other than those directly involved in its use. Often the distinction between people "using" the material and those "appraising" the material is difficult to maintain for some techniques (e.g. individual student tryouts, to be discussed below), but the distinction has nevertheless proved useful.

The techniques for collecting critical appraisals overlap to a great extent with the methods of evaluating objectives described earlier. Collections of opinions from experts of all sorts, teachers, parents, students, administrators, authors, etc. can be accomplished by means of questionnaires, checklists, interviews, panels, diaries, Q-sorts, the delphi technique, etc. The criteria against which each of these populations can appraise the materials will vary. Teachers will undoubtedly be concerned with such factors as congruence of content with their own biases or capabilities, practicality of the format, mode, and/or requirements of the instruction, degree of integration with existing curricula, extent of teacher input, flexibility, and so on. Parents may be very concerned with the type of value system implied in the material, currency of content, orientation (i.e. to college bound or vocationally oriented students), sex or racial bias portrayed, and so on. Any or all of this information can bear upon the subsequent

revision of instructual materials especially when external information supports the critical appraisal.

One judgmental data source which has not been tapped in this chapter which deserves attention is the author. The definition of the term author differs somewhat from medium to medium. In the case of the print media, text, or audio tapes, the definition is fairly easy. But with film or videotape the term author is probably closest in meaning to director. The author is often not recognized as a source of revision information but he is in fact a major if not the major one especially at the early stages of the product. The author makes literally thousands of decisions when he embodies the content he intends to teach in a suitable form, decisions concerning sequence, phrasing, orientation, value, difficulty level, and so on.

When textbook authors write their prose they are writing with a particular audience in mind, with a particular standard of difficulty and clarity. As the sentence is written, judgments are being made as to its adequacy in conveying intended meanings, the sophistication of the audience, the contribution of the sentence to the orderly development of the intent of the paragraphs, etc. If the sentence fails to meet these criteria it will be rewritten until the author is satisfied. It should be obvious that estimating the number of these decisions that the authors make as in the thousands is probably quite conservative.

Authors, however, are often only dimly aware of the decision process. Explicit standards are rare and, probably

as a consequence, consistency in decision making is less frequent than would be desirable. Some profit might accrue, therefore, by increasing author awareness of his decisions. Lawson (1973) has constructed questionnaires and checklists which should prove useful in this regard. In one of his questionnaires, authors are querried on whether specific learner objectives are provided, whether entry behaviors are specified, whether provision is made for learners to enter the product at points other than the beginning, whether the format and display are appropriate for the intended population, whether examples and illustrations used are likely to be of interest to the intended population, etc. The effects of such procedures upon authors is unknown at this time and should be the object of future study. The willingness of authors (or more accurately what type of authors would be willing or unwilling) to explicate their decision processes would be very interesting to examine.

It should be noted that the descriptive and critical appraisal techniques described in this section thus far on formative interim evaluation can be used at the formative product stage as well. The difference is primarily one of the size of the "piece" or the closeness of an interim format that is being evaluated to the final product but the principles involved are generally comparable from stage to stage.

↕ double space

Much useful external information at the interim stage can be gathered by using the same criterion measures which will be used at the formative product stage. If an achievement test is carefully constructed to measure the complete set of objectives of the instruction, there is no reason why, if the test has been carefully criterion referenced, that appropriate items from that test could not be used to evaluate "pieces" of the instruction designed to teach certain of the objectives. If, however, the criterion test does not measure every objective but merely samples from among many, then it would not be appropriate to use that test as an interim evaluative device. Although very desirable at the formative product stage it is mandatory at the interim stage that some evaluative information be provided on every objective of the instructional product.

The principles of construction of and the theoretical bases for external evaluation devices of many types should be familiar to the readers of this chapter and do not require reiteration here. That the evaluation of instructional products should emphasize the attainment of the particular objectives of that product (be criterion referenced) rather than individual differences among students (be norm referenced) is almost at the status of a truism these days. Likewise it is widely acknowledged that evaluation should be as direct and performance based as possible; that is, for example, if students are supposed to be able to correctly assemble an automobile distributor after instruction, they should be

tested by giving them a disassembled automobile distribution, not by testing with paper and pencil their knowledge of the functions of the distributor. Discussion of these issues can be found in any competent educational measurement text but we wish to recommend especially the Handbook of Formative and Summative Evaluation by Bloom, Hastings and Maddus (1971). We will have more to say about more formal external evaluation devices and procedures in the last section of this chapter.

At the interim stage of product development, one should not limit his information gathering activity to highly structured procedures. Much useful information can be gathered in informal types of operations. One which has received increasing attention during recent years is variously called developmental testing (Markel, 1967), individual student tryout (Scott & Yelon, 1969), and oral problem solving (Cunningham, in press). Essentially this technique consists of placing the author (or his agent) with one or more students as they use the materials. Ideally the student(s) will, by means of oral or written comments, help the author locate ambiguities, errors of sequence, and the like, and allow the author to test his assumptions concerning the mental operations which will be employed by students using the material. The students are generally told to "think aloud" as they work through the materials, a procedure which it is hoped will give the author insights into the students' thinking processes and into how well his materials have coordinated themselves with those processes.

Unfortunately, very little empirical knowledge exists concerning individual student tryouts. Beyond an unpublished master's thesis by Robeck (1965) and some recent work by Abedor (1972) discussed earlier in this chapter, very little research on the technique has been completed. The present state of the art is crude, consisting of a number of insubstantiated "tips" as to how to carry off the procedures. And any inspection of the literature relevant to this topic quickly reveals the inconsistency and lack of agreement among those "tips." Some recommend that high ability students be used, others recommend low ability. Some sources argue that students can only clean up semantic and syntactic errors while others insist that the student can make more substantive suggestions concerning sequence, intended prerequisites, etc. Recommendations also vary with respect to preferred level of student incentive, author behavior in the tryout situation, number of cycles of tryout and revision, and on and on. At present few standard procedures can be recommended with confidence. Even the simplest of experiments comparing the quality of instructual products which have and have not used individual student tryouts as part of the development has yet, to our knowledge, to be completed. We hope that, in the next few years, the research necessary to validate and refine these techniques will be completed. A more detailed discussion of the issues and considerations of individual student tryouts can be found in Markle (1967) and Scott and Yelon (1969).

↕ double space

Systematic assessment of context at the interim stage is likely to be wasteful since only pieces of the product are available. The impact of small pieces of a product on a particular context is likely to be unrepresentative of the impact when the product as a whole is integrated into a particular situation.

The evaluator must be aware of intended contexts, however, to guide his choice of students for individual student tryouts or small scale field tests or to guide in the choice of people to conduct critical appraisals. The systematic testing of context and the search for relationship between context and other information about the product is best delayed, however, until the formative product stage.

ADDITIONAL REFERENCES:[5] Materials analysis guidelines (EPIE, 1972), PERT (Cook, 1966), PPBS (McCullough, 1966), systems analysis (Cleland and King, 1968; Kershaw and McKean, 1959).

TECHNIQUES AND PROCEDURES FOR FORMATIVE PRODUCT EVALUATION

At this point in the product development sequence, a version of the complete product is produced. Rather than being discrete, this stage is continuous with evaluation of interim stages of the product. Most often the first evaluative information collected concerning the product as a whole is the same information collected at the interim stage and many of the same techniques are applicable. However, it is the

view here that the major thrust of the formative product evaluation effort should be toward the eventual establishment of the relationship between contextual and other product characteristics. External validity becomes crucial for formative product evaluation activities.

The major techniques for collecting internal information (checklists for descriptive information, questionnaires, interviews, etc. for critical appraisals) have already been discussed and are essentially the same for this stage.

It is also possible when collecting external information to use many of the same methods and procedures as were used in the interim stage including individual student tryouts now with the complete product. The emphasis, however, now shifts to large scale tryout, where the complete product is tried out under the circumstances in which it is supposed to operate. Although having an author hovering over a student is acceptable during developmental testing of an instructional program, it would not be acceptable in a field test of the product.

An inventory of the possible measures which could be collected would be very large indeed, but Metfessel and Michael (1967) have made a useful beginning. They list five major categories of what is here called external information:

1. Indicators of status or change in Cognitive and Affective Behaviors of students in terms of standardized measures and scales.
2. Indicators of status or change in Cognitive and

Affective Behaviors of students by informal or some formal teacher-made instruments or devices.

3. Indicators of status or change in student behaviors other than those measured by tests, inventories, and observation scales in relation to the task of evaluating objectives of school programs.
4. Indicators of status or change in Cognitive and Affective Behaviors of teachers and other school personnel in relation to the evaluation of school programs.
5. Indicators of community behaviors in relation to the evaluation of school programs.

Under these five headings are listed many particular information sources including unobtrusive sources. The strategy we wish to emphasize here is the use of multiple criterion measures in which all criterion measures are recognized as fallible and in need of collaboration by other methods whose fallibilities are likely to be different from the first measure. An attitude scale which purports to measure attitude toward a subject would be more credible if it could be shown to correlate highly with some unobtrusive measure like the proportion of books checked out of the library on that subject or with a classroom observation schedule which demonstrates a high proportion of activities related to the subject during free periods.

Due to space limitation, we cannot possibly discuss all of the many possible sources of external information. Conse-

quently we have chosen to discuss those problems found to be particularly apparent in several product development efforts. At the end of this section we will list references for other major methods for gathering external information.

By far the most frequently sought after outcomes from instruction are student cognitive outcomes, especially higher order cognitive outcomes such as concept learning or problem solving. Yet it often is the case that the criterion measures of these objectives do not allow the inference that higher order outcomes have occurred. Consider the following paragraph which might be taken from an introductory measurement text and some potential test items.

> The mean is the average score of a set of scores and is computed by dividing the sum of all the scores obtained on the test by the number of the scores. If ten students score 1, 3, 4, 4, 5, 5, 6, 7, 7, and 7 respectively, then the mean would be 49 ÷ by 10 or 4.9. The mean is the most frequently used measure of central tendency.

1. The _____ is the most frequently used measure of central tendency.
2. Define mean.
3. What is the mean of the following set of scores?

   1, 3, 4, 4, 5, 5, 6, 7, 7, 7

   a. 4

   b. 4.9

   c. 5.1

4. The measure that is most often used to describe the average score of a distribution is the _____.

5. In your own words, define the mean.

6. Compute the mean of this set of scores.

   12, 19, 15, 30, 57

   a. 26.6

   b. 19

   c. 21.5

Hopefully you will have noted that items 1-3 demand nothing more than verbatim recall or recognition on the part of the student. The student need only remember the form of the information as it was stated in instruction since the wording of or examples used in the test items does not differ substantially from the wording and examples used in instruction. Students answering questions 1-3 correctly could have an understanding of the concept of mean as expressed in the brief passage but the items used to test the concept do not unambiguously allow that inference. Items 4-6 adequately test the higher order objectives in that students probably could not answer those items on the basis of verbatim recall or recognition alone. Key sentences from instruction have been paraphrased, examples have been changed.

As obvious as this point may seem to some, it is apparently not obvious to many of the formative and summative evaluators. Perhaps under the influence of the

performance contracting fad, all too often a very trivial sort of "teaching for the test" can be seen in many product evaluations. Such information can be grossly misleading concerning the actual level of attainment of particular outcomes. Particularly lucid discussions of these issues can be found in Bormuth (1969) and Anderson (1972).

A second problem area is in the procedures for collecting information about effects of the product. Data concerning the effects of a product are not collected at random but, rather, according to some plan which will allow an assessment of the effects of the product in relation to some other state of affairs. It is at this stage of product development where experimental and quasi-experimental designs are useful. The standard reference on this topic continues to be Campbell and Stanley (1963), a truly outstanding summary of the relevant consideration in experimental designs. This chapter should be part of the arsenal of every formative evaluator.

The choice of design for a formative product evaluation is a complicated decision depending upon a number of considerations: cost, utility, practicality, tolerance for certain forms of invalidity, extent of generalizability desired, and so forth. Campbell and Stanley (1963) have discussed the major considerations in the choice of a design: internal and external validity, or, alternatively, replicability and generalizability. The evaluator needs to be concerned with replicability in that if the effect of his product cannot be realiably established, then, of course, decisions about how to make the product better are

meaningless. The formative evaluator must also be sensitive to the extent and type of generalizability of his product. He may not be interested in making generalizations from his evaluation to other products or other contexts than the intended one but within the intended contexts he has to take steps to ensure generalizability. Campbell and Stanley (1963) list eight potential sources of internal invalidity and four potential sources of external invalidity. Each design discussed in their chapter is evaluated against these threats to validity at the consumer of these designs is able therefore to choose those designs which minimizes threats of most concern.

Without doubt, the most frequently used design in product evaluations is the single group pre-test post-test design. In this quasi experimental design a single group of students is first tested to determine how much of the terminal behavior they possess, then are administered the product, then tested again, often with the same test. If learning gains are demonstrated, the product developer will conclude he has a successful product. The problem with such a design is that it allows so many other plausible rival explanations for the observed result: other events occuring between the first and second testing may have caused the results, the pre-test alone may have influenced the post-test, shifts in standards and scoring pre-test and post-test could occur, just to mention a few. Markle (1967) has pointed out that improvements in post-test performance can often be shown to be due to an increased familiarity with terminology used in the product rather than any new learning.

In sum, this design does not have a great deal to offer except that it is probably better than nothing. As the only type of evaluation for the product it is inadequate but as a first step in a more elaborate set of procedures, it can serve a useful function. When it is the only possible design, care should be taken to investigate as many as possible of the potential sources of invalidity specified by Campbell and Stanley (1963). More fruitful designs have been discussed by Glass in Worthen and Sanders (1973).

With regard to contextual information, developers typically have an average student, a particular average classroom setting in mind when they construct a product. It is the function of the formative evaluator to identify and make explicit those assumptions and then to provide a context (if one exists) for the field test. This description implies a two stage process; the identification of intended contexts and then the testing of products within specified contexts. Testing may force modification of the intended context or of the product so that it better fits a more realistic set of context variables. There is thus no mystery to the collection of contextual information in that it involves the use of instruments already discussed in this chapter. Questionnaires or interviews with the author could be used to identify intended contexts on such variables as entering behaviors, student attitudes, socioeconomic status, student interest, teacher experience, teaching style or personality, etc. The intended curricular context for the product including the type(s) of concurrent course work, availability

of instructural aids, and the like should also be assessed. Other variables are discussed in Cunningham (in press) and Sanders and Cunningham (1973).

The identification of actual contexts will center upon the intended contextual variables but the evaluator should be aware of and sensitive to other context variables which might conceivably influence the outcome of the field test. Systematic observation and survey instruments similar to those used for needs assessment could be used to collect this information.

It was stated earlier that the proper focus of formative product evaluation is on the establishment of relationships between context, other information about the product. It was also stated earlier that the purpose of formative evaluation is to provide information of use to the developer of the product concerning potential revisions. These revisions will be most efficiently and effectively made if all of the information discussed in this chapter are available to the evaluator. If, for instance, it is demonstrated that students have failed to master a particular objective, the formative evaluator must find out why and determine what to do about it. He should at that point begin to hypothesize various patterns of relationship among all of the information already collected. Were student entry behaviors over-estimated (context)? Was the read-ability level of the text at that point too great (descriptive)? Did subject matter experts predict difficulty with those concepts and, if so, why (critical appraisal)?

Any one or combination of these considerations could conceivably shed light upon the particular deficiency identified and perhaps imply the steps which should be taken to remedy the situation. In the course of this procedure, relationships are identified which might have some degree of generalizability to other problems within the same product or perhaps even with other products as well.

The potential usefulness of such a focus has been discussed already in Sanders and Cunningham (1973) with respect to a field test conducted by Anderson (1969) in which he found that a discrepancy between an intended context factor and an actual one could account for some disturbing external information. Other examples could be cited. A student of the second author was at a loss to explain why students didn't seem to profit from being provided with knowledge of results in her self-instructional program. A cursory glance at the internal characteristics revealed a great proportion of formal prompts, so many in fact that the program was too easy. Knowledge of results after each frame was simply not needed since more than sufficient information about the correct answer was contained in the frame itself.

To conclude this section, we might reiterate that we recommended the collection of multiple measure of many types of information and the search for relationships among these data. We are not so naive, however, to expect that formative evaluators have unlimited time and resources available to them to pursue all of the recommendations which have been made.

What is proposed is an ideal, a goal to strive for rather than a dogmatic set of prescriptions.

ADDITIONAL REFERENCES:[5] cost analysis (Prest and Turvey, 1965 Tanner, 1971; Fisher, 1971), GFE (Scriven, 1972), ATI (Cronbach and Snow, 1969).

FOOTNOTES

[1] The first author is currently a Senior Research Associate at the Northwest Regional Educational Laboratory.

[2] Information about sampling plans used by NAEP and ISR is contained in several documents published by those institutions. The reader is directed to the list of references at the end of this chapter for references to documents which contain brief summaries of the sampling plans.

[3] A computer program useful for analyzing Q-sort data has been prepared by Bauman (1969).

[4] This example was taken from Gagne (1970).

[5] "ADDITIONAL REFERENCES" provided in this chapter comprise works that have not been previously referenced in Sanders and Cunningham (1973) or selected for elaboration here. References contained in the Sanders and Cunningham (1973) article have not, for the most part, been repeated in this chapter. Hence, the reader may wish to combine the two documents for a more complete treatment of methods for formative evaluation in product development.

[6] The Lawson (1973) article has been written within the Sanders and Cunningham (1973) framework.

[7] This example was adapted from an illustration provided by Berelson (1952).

REFERENCES

1. Abedor, A. Second draft technology. Bulletin of the School of Education, Indiana University, 1972, 48, 9-43.

2. Anderson, R. C. The comparative field experiment: An illustration from high school biology. Proceedings of the 1968 Invitational Conference of Testing Problems. Princeton, N. J.: Educational Testing Service, 1969, Pp. 3-30.

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

4. Bauman, D. J. Computer program for processing Q-sort data. Educational and Psychological Measurement, 1970, 56, 67-74.

5. Berelson, B. Content analysis in communication research. New York: The Free Press of Glencoe, 1952.

6. Berelson, B. Content analysis. In Handbook of social psychology. Cambridge, Mass.: Addison-Wesley, 1954, Pp. 458-518.

7. Bloom, B. S., Hastings, J. T., & Madaus, G. F. Handbook on formative and summative evaluation of learning. New York: McGraw Hill, 1971.

8. Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

9. Budd, R. W., Thorp, R. K., & Donohew, Lewis. Content analysis of communication. New York: Macmillan, 1967.

10. Burnett, J. Ceremonies, rites, and economy in the student system of an American high school. Human Organizations, 1968, 28, 1-10.

11. Burnett, J. Event description and analysis in the micro-enthnography of urban classrooms. Urbana, Illinois: University of Illinois, 1969.

12. Campbell, D. T., & Stanley, J. E. Experimental and quasi-experimental designs for research in teaching. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally, 1963.

13. Clelland, D. and King, W. Systems analysis and project management. New York: McGraw-Hill, 1968.

14. Cook, D. Program evaluation and review techniques, applications in education. U. S. Office of Education Cooperative Research Monograph, No. 17, OE-12024. Washington, D. C.: USOE, 1966.

15. Cronbach, L. and Snow, R. Final report: Individual differences in learning ability as a function of instructional variables. Palo Alto, California: Stanford University, 1969.

16. Crothers, E. J. The psycholinguistic structure of knowledge. Studies in mathematical learning theory and psycho-linguistics. Boulder: University of Colorado, November, 1970.

17. Cunningham, D. J. Evaluation of replicable forms of instruction. AV Communication Review, in press.

18. Davies, I. Task analysis: Some process and content concerns. Audio-Visual Communication Review, 1973, 21, 73-85.

19. Edmonson, J. B. et al. The textbook in American education. Thirtieth Yearbook of the National Society for the Study of Education. Part II. Bloomington, Illinois: Public School Company, 1931.

20. Educational Products Information Exchange (EPIE). Early childhood education, how to select and evaluate materials. EPIE Report No. 42, 1972.

21. Fisher, G. H. Cost considerations in systems analysis. New York: American Elsevier Publishing Co., 1971.

22. Gagne, R. M. The conditions of learning (2nd Ed.) New York: Holt, Rinehart, and Winston, 1970.

23. Grobman, H. Content analysis as a tool in formative and summative evaluation. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April, 1972.

24. Guttentag, M. Social change in a school: A computer content analysis of administrative notices. Journal of School Psychology, 1971, 9, 191-199.

25. Helmer, O. Analysis of the future: The delphi method. Santa Monica, California: Rand Corporation, 1967.

26. Herriott, R. Survey research method. In R. E. Ebel (Ed.) Encyclopedia of Educational Research. New York: Macmillan, 1969.

27. Holsti, O. Content analysis for the social sciences and humanities. Reading, Mass.: Addison-Wesley, 1969.

28. Institute for Social Research. Interviewer's Manual. Ann Arbor, Michigan: Institute for Social Research, University of Michigan, 1969.

29. Kerlinger, F. Foundations of behavioral research (2nd Ed.). New York: Holt, Rinehart, and Winston, 1973.

30. Kershaw, J. and McKean, R. Systems analysis and education. Santa Monica, California: Rand Corporation, Memorandum RM-2473-FF, 1959.

31. Lawson, T. E. Formative instructional product evaluation instruments. Urbana, Illinois: Center for Instructional Research and Curriculum Evaluation, 1972. Also in Educational Technology, 1973, 42-44.

32. Markle, S. M. Empirical testing of programs. In P. C. Lange (Ed.), Programmed instruction. Sixty-sixth yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1967.

33. McCullough, J. Cost analysis for planning--programming--budgeting cost-benefit studies. Santa Monica, California: Rand Corporation, 1966.

34. Metfessel, N. S., & Michael, W. B. A paradigm involving multiple criterion measures for the evaluation of the effectiveness of school programs, Educational and Psychological Measurement, 1967, 27, 931-943.

35. National Assessment of Educational Progress. National Results, Science. Denver, Colorado: Education Commission of the States, 1970.

36. National Study of School Evaluation. Evaluative criteria, secondary school. Arlington, Virginia: National Study of School Evaluation, 1969.

37. National Study of School Evaluation. Evaluative criteria, junior high school/middle school. Arlington, Virginia: National Study of School Evaluation, 1970.

38. Oppenheim, A. Questionnaire design and attitude measurement. New York: Basic Books, 1966.

39. Osgood, C., Suci, G, and Tannenbaum, P. The measurement of meaning. Urbana: University of Illinois Press, 1957.

40. Potter, D., Sharpe, K., Hendee, J. and Clark, R. Questionnaires for research: An annotated bibliography on design construction, and use. Portland, Oregon: Pacific Northwest Forest and Range Experiment Station, 1972.

41. Prest, A. and Turvey, R. Cost-benefit analysis: A survey. The Economic Journal, 1965, 75, 683-735.

42. Robeck, M. A study of the revision process in programmed instruction. Unpublished Master's Thesis. University of California, Los Angeles, 1965.

43. Sanders, J. and Cunningham, D. A structure for formative evaluation in product development. Review of Educational Research, 1973, 43, 217-236.

44. Scott, R. O., & Yelon, S. L. The student as co-author--The first step in formative evaluation. Educational Technology, 1969, 9, No. 10, 76-78.

45. Scriven, M. The methodology of evaluation. In R. E. Stake (Ed.), AERA monograph series on curriculum evaluation. No. 1. Chicago: Rand McNally, 1967. Also, in Worthen, B. and Sanders, J. Educational Evaluation: Theory and Practice. Worthington, Ohio: Charles A. Jones, 1973.

46. Scriven, M. Prose and cons about goal-free evaluation. Evaluation Comment, 1972, 3, 1-4.

47. Simon, A. and Boyer, E. (Eds.). Mirrors for behavior: An anthology of classroom observation instruments. Philadelphia: Research for Better Schools, 1970.

48. Stake, R. Objectives priorities, and other judgment data. Review of Educational Research, 1970, 40, 181-212.

49. Stephenson, W. The Study of behavior: Q-technique and its methodology. Chicago: University of Chicago Press, 1953.

50. Tanner, K. A heuristic approach to program cost/effectiveness analysis. Paper presented at the Annual Meeting of the American Educational Research Association, New York City, 1971.

51. Thiagarajan, S., Semmel, M. and Semmel, D. Sourcebook on instructional development for training teachers of exceptional children. Bloomington, Ind.: Center for Innovation in Teaching the Handicapped, Indiana University, 1973.

52. Torgerson, W. Theory and Methods of Scaling. New York: Wiley and Sons, Inc., 1958.

53. Wolcott, H. An ethnographic approach to the study of school administrators. Human Organization, 1970, 29, 115-122.

54. Worthen, B. and Sanders, J. Educational Evaluation: Theory and practice. Worthington, Ohio: Charles A Jones, 1973.