



The Why and How of Evaluation: *Youth Films Evaluation Project*¹

Ted Ward

Introduction

In this era of technology, evaluation is socially respectable—even expected. From the Consumers' Union to Ralph Nader, evaluation is becoming part of the way of life. Evaluation—especially testing of students—has long been a standard part of formal education. Accountability has become a watchword. Is it worth what it costs? Does it do what it is supposed to? Is X better than Y? But are the new questions being asked within the educational establishment, replacing or augmenting the historical evaluative questions focused on “how well is Freddy learning?”

It is to be expected that the producers and supporters of the Youth Films (and the users, of course) are interested in evaluation. Regardless of the degree and extent of good intentions and well-meaning actions, the issue remains: how well are the films doing their job? It is similarly logical that specialists in evaluation would be able to get an answer to this question.

Research of any sort is generally limited by the questions it is able to ask with precision. The findings of an evaluation can be no more precise than the questions posed. For the present study the questions were limited to those that could be asked during the three months remaining in the school year (March, April, May). Longitudinal questions, such as those dealing with long-term impact and with case study effects could not be raised. The present research was also limited by the fact that the assumed question (“are the films doing their job?”) could not be asked in any precise manner since only relatively generalized statements of objectives were available for the individual films and for the program as a whole. Thus the first task was to identify worthy questions which could be answered within the time and cost constraints of the requested study.

A. The Basis Of Evaluation

In keeping with the whole of educational technology, educational evaluation has been refined and developed extensively in recent years. Today evaluation is divided into two species or classes according to its purpose. Two distinct species of evaluation have been defined in the professional literature: *formative* and *summative*. Since these are distinctions related to purpose, they do not necessarily indicate separate kinds of data gathering or different kinds of concerns about reliability and validity. *Formative* evaluation is intended to provide information on which significant improvements can be made in an instructional system or component. *Summative* evaluation is intended to provide information on which judgments can be made about the effectiveness and worth of the educational experience and the materials and procedures that produced it.

Most competent evaluations are based on a series of four operations best described as a series of steps or stages: the base or first stage is *description*. On this is built *measurement*, on measurements are built *assessments*, and then by bringing value positions to bear on the assessments one can make *evaluations*. Usually descriptions are verbal, though graphic and pictorial descriptions are often more meaningful.

¹Ted Ward. The Why and How of Evaluation: Youth Films Evaluation Project, July 1973

The *measurement* step is necessary if clear-cut assessment is to follow. Most scientists in the physical and social science domains will insist that descriptions that have not been quantified are an inadequate basis for comparative analysis. That differences may exist can sometimes be determined “by inspection,” but without *measurements* (expressions of the descriptions in numerical terms) the significance of the contrasts can be stated only subjectively.

Comparison of two or more measurements constitutes *assessment*. Two scores representing two different learners can be used to assess that one has achieved more than the other. Two measurements taken on the same learner at two different times can be used to assess that a learning gain has taken place. Even the statistical significance of these differences, gains, losses, relative competencies, etc., can be determined as assessment activities. But to say that the gain is *important*, that the learner has achieved *well*, or that the competency is *adequate* requires a step beyond assessment; these judgments are representations of the outcomes of *evaluation*. Evaluation is the process of putting a value judgment on the conclusions of an assessment. Or, putting it another way, evaluation is judgment of the worth or importance of an empirical finding.

The relativity of “goodness” in educational evaluation is illustrated by the contrast between conclusions one draws from two different value positions: assuming that educators are concerned about both *growth* and *competency*, each can be taken as a value position. Assume that the following scores for a set of five learners represent a pre-test before a learning experience. (The students are named Alice, Bob, Chuck, Donna, and Ed.)

$$A=4, B=12, C=10, D=7, E=4$$

After the learning experience, the following scores might result from a second measurement on the same test used for the pre-test:

$$A=18, B=14, C=14, D=14, E=9$$

An assessment of these measurements allows expression of the scores as difference, “growth,” or learning scores:

$$A=+4, B=+2, C=+4, D=+7, E=+5$$

Which learner did *better*? Who is the *best* or who is *best* cannot be answered until we take a value position. For example, if one assumes that competency is the value (“competency is more important than growth”) then clearly Alice is more competent in this skill or knowledge—before and after the learning experience. Ed remains the least competent. Bob, Chuck, and Donna seem now to be equally competent. This evaluation is based on an assessment of the differences of the various post-test scores in light of the *competency* value position. Thus when we note that Alice “did better” we are evaluating on competency.

If we accept another value position, that growth is more important than competency, then Donna “did better.” In terms of growth, Donna’s 7-point gain is the largest, and for this value position the largeness of gain is the basis of judgment or evaluation. Note that this evaluation uses assessment data from the pre-tests and post-tests expressed as amount of differences between the two.

B. Special Problems of Evaluation of Instructional Materials and Procedures

In evaluating an instructional material or procedure, we are sometimes concerned more about growth than about competency. This is especially true when public and private elementary and secondary schools are the focus or locus of evaluation and when no common agreement exists about what constitutes competency in the given subject matter.

The lack of standards or norms in a given subject or area of learning increases the difficulty of *summative* evaluation, especially. Summative statements in such a case are necessarily limited to comparisons (material X “does more’ than material”) and to pragmatic statements about whether or not an instructional material “works” (has effects on learners). Evaluating whether the effects it has are “adequate” to bring a person up to norms and whether the rate of change is worthwhile, in terms of cost-benefit ratio or in terms of instructional time expended, are difficult, if not, impossible tasks.

When criterion problems are hard to resolve, summative evaluation is often diverted to focus on other and less worthy questions than the worth of the learning gained. For example, it is common to find summative evaluations concentrating on the teachers’ opinions or estimates of the worth of the experience, the students’ liking or disliking of the experiences and materials as reported by opinion surveys, and the use of highly subjective and non-representative anecdotes as a basis for passing judgments. This sort of evaluation is potentially misleading unless kept in its place within a more comprehensive evaluative framework.

Measurement problems also thwart well-intentioned evaluation. Especially in the affective domain, where learning is very closely related to life-style-determining personality and value-system variables, measurement is extremely difficult. Verbal responses in the affective domain often bear little resemblance to the concomitant behaviors. (“What you *are* speaks so loudly I cannot hear what you say.”) Verbal data on written tests or interviews can rarely be taken at face value. It is too easy to misrepresent oneself, enhancing one’s ego or to suppress guilt and fears; and one may not be sure enough of his or her own feelings (“the real me”) to be able to speak or write with confidence. Since unobtrusive observation of the behaviors that reflect true affect are awkward and usually very costly, verbal self-report is usually utilized to get quantifiable data. The validity of such data is always open to challenge.

With the advent of well-specified behavioral objectives as the popular way to begin the design of an instructional system or component, another problem of evaluation has arisen. One of the overly-ambitious assumptions that tends to accompany the emphasis on behavioral objectives is that because a learning objective is well-defined it is likely to be achieved.² The tragedy of this assumption is that one tends to seriously over-expect results from a given learning experience. Behavioral objectives have a characteristic tone of finality. For example, “When confronted by the question, ‘What is the sum of two plus two?’ the learner will respond ‘four,’” has the ring of *always* and *forever*. Experienced educators are more familiar with the tentative nature of new learnings and the distinct possibility that a learner will know something today and not know it tomorrow. The *persistence* of learning is not the issue here. The tendency to ask evaluative questions that are too ambitious (too demanding) *is* our concern. The probability of any given brief instructional experience leading directly to life-changing

²This assumption is reflected by the technologist’s views that the objective should serve as the basis for an item in the achievement test and that a well-designed instructional system is based on the realizable ideal of all appropriate learners achieving all specified learning objectives.

consequences is remote. Thus it seems unfair if not irresponsible to take a list of intentions ascribed to a given instructional material and to evaluate an instructional material exclusively on the basis of pre-test and post-test of these traits or behaviors in learners.

C. Selecting the Questions to Guide the Summative Evaluation

For the films in the Youth Films program no lists of specified instructional objectives are available, yet there are strongly worded intentions reflecting a set of fervently held values and aspirations. It is only reasonable to expect that one viewing of any one film is likely to be, for most learners, just one more media-based message in the mosaic of experiences that constitute the media-saturated life of today's American. Nevertheless there are useful questions that can lead to summative evaluation of certain important aspects of a given film. For example, if a value position is held that a film should be *meaningful*, then it is possible to evaluate on the basis of an assessment comparing intended message and message received. If the capability of a film to hold the *interest* of the learner is valued, an evaluation can be based on an assessment comparing immediate and long-term recall or comparing the intended and realized responsiveness of a viewer or group of viewers.³

In strictest terms, a *descriptive research* such as pursued to examine the sample of films for Youth Films, is not the most explicit form of evaluative study. The constraints of time and costs, as well as the measurement problems associated with the affective domain, have dictated that a substantial part of the study utilize descriptive research methodology rather than purely evaluative methodology. Thus the report is more focused on what seems to be happening than on how good or how important it is.

The questions that we sought to answer through the descriptive research were the following:

- 1) Does the film carry a message that is meaningful to the intended learners?
- 2) Does the film tend to nudge the learner to examine values positions and rationales that are higher than those he or she currently holds?
- 3) Does the film tend to affect the learner's basis of behavior, and, if so, in a desirable direction?
- 4) Does the film fit into a supportive educational environment, both in terms of the physical arrangements and the emotional climate within which the films are used?

These questions are summative in nature and their answers, in the affirmative, would seem to be basic to any claims for satisfactory performance of any film in the Youth Films program.

D. Selecting the Questions to Guide the Formative Evaluation

Formative evaluation is somewhat easier to describe, yet the questions it suggests are so abundant as to make an exhaustive list unworkable. Virtually any information that sheds light on ways to improve an educational experience can qualify as formative. The issue for designers of evaluation is to assure responsible validity and reliability of such information. Theory becomes especially important at this point. Value positions must have clear relationships with the instructional experiences themselves. What is valued as the learning outcomes must be synchronous with the nature of the instructional experiences. Assessments upon which evaluations will be based will have to compare experiences induced by the material and specifications of experiences that are theoretically contributory to the in-

³ Attention is directed to the superiority of evaluations based on this sort of direct investigation of the target audience over the indirect estimation by an intermediary person or group, e.g., asking teachers if they think the film was meaningful for their students or if it was interesting to the students.

tended outcomes. Thus the key question for a formative evaluation of an instructional material or procedure is the following:

To what extent are the observed experiences resulting from the instructional material or procedure consistent with the theoretical model of the appropriate learning experience?

From this question two others are derived, and thus are established two pursuant lines of inquiry:

In what ways is the instructional material or procedure most deficient in comparison with the appropriate learning experiences for the intended learning outcomes?

What suggestions about improvement of the instructional material or procedure can be deduced from descriptions of their uses?

The first of these two derived questions is more truly within the domain of evaluation since it can be answered by comparisons of data against a value position. The second question is hardly recognizable as a classical focus of educational evaluation, but it is the crux of formative evaluation. It leads to the gathering of descriptive data about level of motivation and learning gains at several points in instructional experience. The particular data collected and the uses made of the findings are subservient to the creative process of improvement of the material or procedure.

Formative evaluation is ordinarily associated with the pre-publication or pre-release pilot study of effectiveness of a new instructional material; as such, formative evaluation is expected to lead to changes. After instructional material has been released for sale or large-scale distribution, continuing to carry on formative evaluation makes sense only if a revision or subsequent version is anticipated.

In the case of a continuing series of materials or publications of which there are to be subsequent issues (not revisions of previous releases), there is potential value of formative evaluation that leads to improvement of the series. Here again, theory is especially important. The formative evaluation must be focused on improving the *guidelines* and *models* of instruction upon which the productions are based—somewhat a more abstract task and hence more demanding than the improvement of a given material.