But Will It Work? Formative Evaluation in Foreign Language Materials Development

This article is intended to help consumers of media-based learning products make informed choices. Through an understanding of the history and process of product development consumers are better equipped to ask questions about the comprehension, appeal, user friendliness, and persuasiveness of learning products—print or electronic—they are about to recommend for adoption or purchase.

What is Formative Evaluation?

The term "formative evaluation" refers to procedures which can be used to help determine—during design, production, and evaluation—whether a particular media-based product will produce specific outcomes with particular audiences in a predictable fashion. Guided by the scientific method, formative evaluation is particularly concerned with validity and reliability. A method is considered valid when it teaches or produces what it intends to produce or teach; it is reliable when, if used under different conditions within a defined range of parameters, it produces similar outcomes. Although no one can say with accuracy or absolute certainty that something is going to work, we can significantly increase the probability of success with the procedures known as formative evaluation.

It might come as a surprise to many of us in education that those who produce the textbooks, videos, and software for the teaching and learning of foreign and second languages (this includes classroom teachers turned producers) do not use formative evaluation consistently or very well; as a matter of fact, commercial television and corporate producers use formative evaluation more effectively and systematically in producing entertainment and training materials than do the producers of the teaching materials we use in our language classrooms. In part not using formative evaluation may be due to ignorance on the part of producers; in part it may be a lack of understanding of how to plan for the testing of the materials. Whatever the reason, we, as consumers, may also be partially to blame because we are not informed about formative evaluation either; we may not know what is involved in formative evaluation or what questions to ask developers of classroom materials—issues and questions that should have been addressed during the development of the materials. In the Age of Information, replete with easily accessible technologies, formative evaluation is no longer a question of **if** it should be done, but rather a question of **when** it should be done.

Background

For many years, the textbook was the primary and preferred tool—among a relatively limited number of learning tools—of teachers (to help structure their pedagogy) and of students (to help serve as reference guide and source of learning activities). Many textbooks were written by highly regarded and esteemed teachers for use by other teachers. In the past—especially in the United States—teachers were the principal sources and arbiters of academic learning experiences. Today, learning experiences have the potential of being much more individualized: The availability of computers, videocassette recorders (VCRs), and cable television stations appeal directly to learners of all ages, tempting and encouraging them to manage their own language learning.

As more and more learners succumb to piloting their own learning, instructional materials will have to "stand on their own" with learners; the increasingly popular group of stand-alone materials include computer-based instruction, audiotapes, videotapes, and computer-controlled videodisc instruction.

Formative Evaluation: A Key to Profits in the Commercial Marketplace

As we begin to consider the issues involved in formative evaluation, it is helpful to look at how people who must produce predictable outcomes cope with formative evaluation. We, as foreign language professionals, may feel little or no kinship with producers in the world of advertising; what happens there in terms of formative evaluation is instructive, however, and can offer some startling insights.

Even the best teachers among us have a difficult time competing with the premiere teacher of today's world, namely the insipid television commercial. Why? Through very careful study of perceptual psychology and physiological reactions of target viewers, producers of television commercials create very powerful messages that reach and teach not only children but also incredible numbers of adults. One 30-second commercial may involve hundreds of hours of dogged research with human subjects under carefully controlled conditions. (Woog, 1988)

In formulating a typical Fall line-up of prime-time shows, a commercial network spends millions on pilots (a sample episode of a proposed series) and gathers data via the Neilson or Arbitron ratings to determine which pilots will go into full production (based on audience response). Although one episode of a typical situation comedy (sitcom) costs only $200,000, networks typically spend five times as much up front to determine whether or not the show will be a "hit." The stakes are high. If the sitcom does not draw the demographic elements—the thousands of potential viewers (consumers) of a sponsor's products—it will, in all likelihood, never be produced, or if episodes have been produced, they will be cancelled.

In fields such as advertising, television, and industry formative evaluation goes by other names, such as "market research," "field testing," "alpha-" and "beta-testing." No matter what the label, the concept is similar. In the corporate environment, the need to provide standardized training for large numbers of employees who have limited or no access to human teachers created the need for valid and reliable learning outcomes; in the television industry, there is a direct relationship between the outcomes produced with target audiences and profits. Although the hard realities of market volatility and stockholder reaction force commercial developers to chart very carefully the course of their product developments (millions of dollars can be lost when a project is improperly planned and tested), the textbook industry is buffered: Standing between it and the reactions and demands of consumers-as-learners are teachers-as-consumers.

Whereas commercial television producers must please the consumer of television programming directly, educational textbooks and media-based learning products producers are most dependent upon teachers' purchasing decisions *even though the student learners are the ultimate consumers of media-based learning products.* Educational products that are accepted or rejected solely on the basis of student reactions are extremely rare.

As the demand and production of modular media-based materials grow, students-as-consumers of such materials will increasingly exercise their acceptance and/or rejection of such products. Like it or not, students as consumers of learning will engage in formative evaluation. For insight into how formative evaluation can be instrumental in the production of media-based learning products that will receive positive user acceptance, let us consider the case of educational television.

The Children's Television Workshop: A Model Case

A successful developer of educational television programs, the Children's Television Workshop (CTW) is an insightful model for examination: Its procedures have been openly documented, and it has devoted itself to finding creative ways of doing formative research specifically for educational purposes. The procedures developed by CTW for formative evaluation are designed to produce appealing and comprehensible learning products; the evaluation procedures are now widely used in commercial television and corporate training productions.

In determining the format and presentation devices for its science series—*3-2-1 Contact*—CTW sought to strike a balance between the theoretical research of educational experts and the practical experience of television production professionals.

To make sure that the main points of view were explored, CTW assembled three teams of experts: The scientific content team assured the accuracy of the material; the formative research team concerned itself with the target audience; and, the production team coordinated and implemented the inputs of the content and research teams in the creation of the final product. (Mielke, 1983)

One of the goals of the series was to make science more appealing to girls. Research seems to confirm that boys generally have a more positive attitude and interest in science. (Mielke, 1983) Formative researchers set out to find what factors—in casting, scripting, and theme—would tend to encourage girls to become involved with the science topics of the program. To isolate possible factors, the formative researchers showed still photographs and programming segments to both girls and boys, asking them to indicate which characters they liked best, which topics they preferred, etc. As a result of this research, a general preference profile for each sex emerged: Boys preferred male cast members; girls preferred female cast members; animal characters—particularly mammals—were favorably perceived by both sexes; girls preferred animal themes to space themes, and boys preferred insect and snake themes. (Mielke, 1983)

Formative evaluation—based on the criteria established by field experts—illustrates the need for integrating both the theoretical and the practical points of view. While the CTW people had a theory about girls' interest in science, formative research revealed ways to tap this interest, and thereby, enhance it. As Mielke suggests, the "stimulus complexity" of television can greatly complicate the translation of principles into programs. It is this translation process that formative evaluation can effectively guide.

The Stages of Formative Evaluation

Any media-based development plan needs to incorporate the five basic stages of formative evaluation: needs assessment, pre-production formative evaluation, production formative evaluation, implementation formative evaluation, and field testing. Each of the five basic stages has four possible areas of inquiry or components. These areas of inquiry or components are *the* issues about which producers, learners, and teachers must ask questions; for the producers of learning products, these four components determine how reliable and valid the product will be; for teachers and learners, these components effect how successfully teachers teach and learners learn using the product. When considering production or purchase of any media-based learning tool, the following components of formative evaluation must be considered: 1) comprehension, 2) appeal, 3) user friendliness, and 4) persuasiveness. (Flagg, 1989)

The Components of Formative Evaluation

Comprehension

Evaluation of any learning product—or pedagogical practice, for that matter—begins with the first and most important component or issue, namely comprehension. Any learning product or practice that fails to convey the information that it intends to convey is flawed. Flaws in the comprehension component directly affect all the other components.

Foreign language teachers and learners frequently encounter the comprehension component in introductory language textbooks in the form of the ubiquitous grammar explanations or the text itself entirely in the target language; in the classroom, the comprehension component shows itself in "the student speaks 90% of the time in the target language." Presumably, the teacher speaks 10% of the time in the target language. Although there are numerous philosophies and shibboleths about the issue of target language use in textbooks and classrooms, how much testing of such philosophies-shibboleths has actually been done? How much do we really know about how learners piece together their knowledge about language? What are our real goals when we write grammar explanations or force learners to speak in a target language that no beginner can realistically understand?

Language teachers and learners alike also encounter the comprehension component in cultural materials which, to be effective, often require changes in attitudes. Although attitude change is an important goal, such a goal will not be reached if the message is poorly presented to the targeted audience. For example, say a producer wants to make a video about life in the French *lycée* for an audience of American high school students. The targeted audience harbors an attitude that it has nothing or very little in common with French teenagers. In spite of this perceived lack of common ground, the producer wants to demonstrate that young people in France and young people in America share many of the same concerns.

To demonstrate common ground, camera shots and angles must focus on what French teenagers are saying and doing and not on the superficial differences of school environments or other distracting details. Furthermore, if the videotape footage is improperly edited, the conversational flow suffers; whereas a natural feel in conversation is desirable, using clips in which French students use many difficult and complex idiomatic expressions hinders the comprehension of American students with beginning French language competence.

In evaluating the comprehension component of a media-based learning product like video, not only producers of such media but also teachers and learners must ask the following: 1) Does the product convey the information intended in a manner comprehensible to the user? 2) Is the message or theme clearly presented? and, 3) Does the information presented clearly support and clarify the message or is it merely a case of information overload, that is, lots of information which has little or nothing to do with the message?

Are there techniques that can be used to test the comprehension component in order to determine how adequately it has been treated in a given learning product? Yes there are, and the most commonly used technique—the cornerstone of educational research—is the pre-test and post-test. Normally, these are paper-and-pencil tests which ask specific content questions, either open-ended or multiple choice. The oral interview is another device or technique which can give clues to how well learners have comprehended the material to be learned. When dealing with video or software, a technique known as "stop-program" interviewing is effective in getting specific information from the learners while they are actually viewing a specific learning program.

Incomprehensible material, confusing messages, and irrelevant information are all part of the comprehension component; unfortunately, they also affect the second component of formative evaluation, namely appeal.

Appeal

Appeal may be defined as the intensity with which learners engage themselves in using learning products. When evaluating a learning product, it is well to begin by asking how closely learners can identify with the material. Identification with the material of a learning product—media-based or text—depends on a number of factors, not the least of which is the degree to which the subject matter itself is interesting, fun, and enjoyable.

A simple technique or method of testing the appeal component is to give learners a sheet of paper with a graduate scale or a yes/no response scale on which they indicate their likes or dislikes of a particular segment or aspect of the learning product during or immediately after its use. Another technique involves simply observing the facial expressions and attentiveness of learners from a strategic location in the room. In addition, open-ended questions—on paper, in an interview, or a discussion group—can elicit information about how appealing a learning product is or is not.

User Friendliness

Of the four components of formative evaluation, user friendly has probably reached the rarified elevation of buzzword status. As an issue or category, this component is particularly critical in interactive media where navigation through a series of visuals or text screens is paramount; it applies, however, to any product or practice touted as instructional. Moreover, to a greater or lesser degree, we expect any reality to be more or less user friendly.

For example, when we buy a piece of clothing, we expect a user friendly tag in the back. While not the tag's only purpose, the tag as standard helps us in getting dressed when we cannot easily distinguish the front from the back. The same holds true for products we use in learning: we expect user friendly standards. With textbooks—when table of contents or indexes are not properly set up or when page layout is confusing—learners can have a difficult time in finding needed information. (Duffy & Waller, 1985) With computers, the user interface is an oft-cited cause of program ineffectiveness, that is to say, program unfriendliness. Many of us are insecure about navigating computer programs; learning every program author's peculiar standards of execution adds insult to injury. Anyone developing interactive media must be aware of the written and unwritten standards of computer screen layout in order to maximize user friendliness.

Techniques that test whether or not interactive media standards work involve looking at actual users of such interactive programs. Ideally, users are observed during their first few "bouts" with the program. An unfriendly program with a poor user interface will usually be mastered by learners; in the meantime, the frustration of figuring out how the program works takes its toll. It is best to catch the factors contributing to poor user interface early in the program developmental process. An effective technique is to ask learners to "think aloud" as they navigate the program for the first time. Direct observation and "thinking aloud" of large groups of learners are often too costly; instead, asking learners to write down their impressions of a program as they use it can yield valuable clues about the program's user friendliness.

Persuasiveness

Just because a given learning product scores high on the components of comprehensibility, appeal, and user friendliness, does that mean it will be effective in motivating learners to modify or change their behavior—evidence that a desirable learning outcome has occurred? Not necessarily. Any learning product designed for instruction and learning must deal with the fourth component of formative evaluation, namely, persuasiveness, in order to be effective in promoting attitude change.

In examining and testing persuasiveness, we encounter special problems, ranging from the moralistic aspects of imposing a particular set of values on learners to the special difficulties of actually measuring attitude change. The various substance abuse programs are typical of the problems society encounters when trying to change attitudes and behavior through educational media-based learning products. The various approaches of foreign language professionals in bridging cultural gaps and weaning students from an addiction to stereotypes are typical of various attempts to change attitudes and behavior in the language classroom.

Of all the components comprising formative evaluation, persuasiveness is the most difficult and frustrating to test. Learners must be asked directly about their feelings regarding the themes or issues of a particular learning program or product immediately after the presentation of the program and then over time. This kind of follow-up testing requires planning; however, the more longitudinal the study, the more we can rely on the conclusions about how persuasive or non-persuasive a particular learning product or practice is in changing attitudes and modifying behavior.

Formative Evaluation in the Production of Foreign Language Teaching Materials: A Successful Case in Point

Presently, it is difficult to find examples of formative evaluation in the development of materials for teaching foreign languages. One rare example is the series, *French in Action.* (See J.E.T.T., Volume XX, No. 2/3, Summer-Fall, 1989, p. 2) The series consists of 52 video segments from which the following were developed: a textbook, a workbook, audio cassettes, a study guide geared toward independent learners, and an instructor's guide for the on-campus course.

At the behest of one of its funders, the development team for the series used an outside consulting agency (Research Communications) to conduct two evaluations: One evaluation was a study of faculty and administrator reactions to a rough-cut of the video for Chapter 14, together with the supporting written and promotional materials. (As purchasers, teachers' opinions counted heavily!); A second evaluation was concerned with student reactions to the video only. The first evaluation was called an "expert evaluation": Teachers were asked how they *thought* the materials would work for their students. The second evaluation is an example of what is known as a "decision-oriented study": Subjects are tested directly to help program developers make tactical decisions about objectives, content, pedagogy, interactivity, and production formats. (Flagg, 1989)

Formative Evaluation and *French in Action:* Testing Appeal

Groups of intermediate French students were tested at colleges in four geographically dispersed locations. The population was broken down into adult students (over 23 years of age) versus traditional students (under 23 years of age); those who had had previous conversational French experience versus those who had not.

During the video viewing, half the subjects rated the video on appeal while the other half rated it on comprehensibility. A test administrator called out the number of each video segment; at the end of each segment, student viewers were asked to mark their rating sheet in terms of whether they liked/understood the segment.

After the viewing, students were given a questionnaire comprised of two types of questions. Fifteen questions queried students about their opinions of various aspects of the program. Sample questions included the following: "I could learn French well from programs like this (Yes/No)" and "Given your level of proficiency, did you find this program to be: Very Difficult—Very Easy". Eight questions were specifically aimed at students' comprehension of the story line. For example, "Mireille is a student 1) American History 2) European History 3) Architecture 4) Do Not Know.

After all students completed the questionnaire, the test administrator led a group discussion of the program segment. The purpose of such a discussion is to evoke—in an unpressured and neutral situation—comments from students that give researchers valuable insights that could otherwise be overlooked. The administrator posed questions such as "What are the strengths/weaknesses of the program?" or "What comments do you have on the use of the story to teach French?" Transcripts of audiotapes of the discussion provided the developmental team with valuable data. The main purpose of testing the appeal of *French in Action* was to gather marketing information. Since this testing occurred fairly late in the development of this series, the testing of appeal would fit into the stage known as implementation evaluation.

Conclusion

Media-based products, moreso than textbooks, must stand on their own with learners, and at the same time, meet the goals of teachers and administrators. Those of us charged with recommending adoption or purchase of learning products—print and electronic—do well to request examination copies prior to help us decide. Over and above that, however, we must ask how the materials were tested during development and whether or not there are any follow-up studies about the comprehensibility, appeal, user friendliness, and persuasiveness of the materials.

When individuals, schools, or publishers undertake learning product development, they must budget time and money for development and implementation of a formative evaluation plan; it is a costly and expensive undertaking. Since foundations that often support the production of learning products are beginning to demand more accountability, those who would produce the learning products of the 1990's and beyond must use formative evaluation to glean valuable data from real learners in real settings. If instructional technology is to regain momentum in the schools, developers of learning materials must be able to provide information about their learning products' effectiveness, validity, and reliability. Schools need reliable information to help them in making decisions about which learning products will best serve the needs of their learners.

References

Flagg, B. (1989). *Improving electronic learning materials through formative evaluation.* (In press). Cambridge, MA: Harvard University Press.

Mielke, K. W. (1983). *Children's understanding of television.* in J. Bryant and D. R. Anderson (eds). New York: Academic Press.

Woog, A. (1988). The meter men. *Profiles, inc.*

J.E.T.T. Contributor Profile

Carolyn Fidelman is a teacher of French, educational media consultant, and interactive technology specialist. She is currently a research associate and director of a FITSE project at Tufts University. Interested readers may write to her at the following address: Arena Annex, Tufts University, Medford, MA.