Past Imperfect, Present Conditional: Some Research on Formative Evaluation

George L. Geis
Ontario Institute for Studies in Education

Fred Keller is an aphorist. He emits oracular statements that can stay with you like a good dinner; and you slowly digest them and return to them for additional nourishment. I recall two of these and, if they were the items on a test of his aphoristic ability, I would suggest that he is batting above .500.

The first one was when I was an instructor in the Introductory Psychology course at Columbia University. After about six months in the job I went to him for some feedback. How were we the staff doing and what improvements could be made? Fred contorted his face and body—as you know he thinks with the large skeletal muscles as well as the small ones—and then he said: “Well, I wish I’d hear you folks say ‘I don’t know’ more often.”

I have returned to that statement again and again as I study instruction and teachers. It says to me that the teacher’s or instructional designer’s role need not be that of savant, of impeccable giver of truth. It suggests a team of two people—teacher and learner—ignorant to only slightly different degrees, both trying to discover things. I would give him 100% for that aphorism.

Because I believe we do not know, I can see our past efforts at instructional design as imperfect and that our present efforts must be put forth conditionally. Each instructional product is tentative; it should be presented conditionally until its validity is demonstrated. And so I became fascinated with, and comfortable with, the topic of formative evaluation—a technique for telling us when and what we don’t know.

Now, about the second aphorism. Seeking ways to get my career started, and I am sure, ways to get me off the rolls of graduate students, Fred suggested that I take a job at Hamilton College in upstate New York. The college had just received a large grant from the Ford Foundation—this was in the ’60s—to produce a number of programmed instruction courses. I was to go there as an expert (never having had any experience in programmed instruction other than having read the Holland and Skinner program). Fred reassured me, saying, “This instructional technology is just an application of what you’ve learned here—operant conditioning and so on.”

I grade him lower on that aphorism for using the word “just.” Either I have been making an awful lot over an awful little during the past twenty-five years, or the development of a valid technology of instruction is more difficult and complicated than he suggested. I choose to take the latter view. (Still, knowing him I add some credits since he was probably trying to allay my fears.)

The so-called “application” has really been a continuing attempt to develop a technology. It draws upon operant psychology but it has selected knowledge and techniques from other specific research-based areas such as text design as well as from the less tough minded areas such as theories of change and innovation. Like all engineering efforts it seeks to do the job better. It reaches out to any likely candidate who may help. If understanding as well ensues, that’s fine, but the technology cannot wait for it. Since most of the technologists are trying to find ways to make things happen and happen better, there are relatively few of us who are privileged to pursue programmatic research. The observations I am making are that the task of applying principles is not a simple one, that the growth of a

This paper was prepared for the Annual Meeting of ABA (Milwaukee, WI, May 27, 1989). Along with many others at that meeting it was prepared for and is dedicated to Fred S. Keller on his 90th birthday. Some of the research cited was sponsored by FCAR (Quebec) grants.

technology is in itself a critical part of continued application, and that there is exciting research being done and to be done.

Formative Evaluation

I have spent part of the past 10 years trying to carry out the kind of research that is appropriate to a technology of instruction in the area of formative evaluation (see, e.g., Geis, 1986, 1987). In conducting formative evaluation one seeks information from either of two sources, experts or potential users. Either set of informants can be brought into the picture at any stage in the design and development of the instructional system. Usually, however, formative evaluation is seen as a particular stage in the development process. Enough of the instructional design should have been planned so that there is a clear picture of the whole and at least representative bits of the instruction itself have been completed in draft form. So the developer, when undertaking formative evaluation, has a sizeable chunk of the instruction in useable form.

Expert as information source. Different *Experts* may be asked to review the materials from different perspectives. Thus, the developers may call upon subject matter experts who can comment on the validity, emphases, and currentness of content, or instructional experts who can review the design characteristics, or specialists in the target population who can comment on appropriateness to audience, or gatekeepers who might offer suggestions and admonitions about cultural or religious acceptability.

Our research team (consisting of my colleague Cynthia Weston at McGill, a fine group of graduate students, and me) is interested in identifying what kinds of information each expert might provide and how the review task can be structured to elicit the maximum amount of useful information from each. As well, we are interested in determining redundancies among experts and between experts and student learners. Look at these concerns this way: if one is interested in

cost-effectiveness (and in a technology that is a primary interest) what would be the best way to use experts as informants to contribute to the improvement of the developing instructional system?

We are carrying out a number of studies of experts. Let me mention one that Alenoush Saroyan (1989) has just completed as her doctoral research. She used two sets of experts, full-time professional instructional designers and microbiology experts who were also teachers. The subjects were asked to review and provide comments on one module in an instructional package which was designed to teach Microbiology to Nursing students. They were instructed to "think aloud" continuously during the review, noting what they were looking at, what problems they saw with it, how it might be changed, and so on. Alenoush then analyzed the transcripts of their verbalizations to reveal the frequencies of certain kinds of statements, and the points at which they noted difficulties and suggested revisions. We hoped as well to see an overall strategy of reviewing.

I realize that, for this audience, the think aloud method (typically associated with the dreaded Cognitive Psychologists) may be as inappropriate as describing pork recipes to rabbinical students or commenting on the efficacy of the Pill during an audience with the Pope.¹ But we did it and the results so far have been very interesting to us. One of the significant findings is that reviewers flag different problems. Although there was some congruence, frequently the indication of a problem with the instruction was unique to a reviewer. That means that, unlike Gertrude Stein's roses, reviewers are not interchangeable, at least reviewers who have not undergone any norming training. Another significant finding is that the strategies of the two types of reviewers were quite different.

¹ I have been forcefully reminded by my friend Susan Markle that think alouds predate Cognitive Psychology and specifically that she was long ago recommending them as a means of collecting feedback from students, as in her excellent guide to programmed instruction (Markle, 1978). One suspects as well that good teachers throughout history

Instructional designers seemed guided by a heuristic. They systematically compared content of the module with its stated objectives, the objectives with test items, and so on. They flipped through the text checking on the congruence of parts. As well, they often played the student, reporting that this or that was likely to be confusing, seemed contradictory, or was unclear. The content experts, on the other hand, proceeded linearly through the text, remarking most frequently on what they saw as factual and terminological problems. They were often critical of surface features of the text, spelling and punctuation, for example. They almost never referred to parts of the module other than the part they were examining at the moment; nor did they comment on the possible problems of learning from the text.

The only formative review, other than by the editor and author, that most commercially published school instructional materials get is by one or two content experts (who may also be teachers). Our findings suggest that although some review is better than none, maximum improvement would require more differentiated sources of feedback. They also suggest that the efficiency and effectiveness of reviewers might be enhanced by training or at least by a good set of guidelines. (In another study, we have examined existing guidelines for reviewers and have published the results of that examination, Saroyan & Geis, 1988.)

Students as information source. The second source of information that I mentioned earlier was the potential student who might be used to indicate the effectiveness of the materials, to flag errors, and even to suggest revision. The literature strongly urges that students be used in diagnostic testing. But: Who?, How?, and When? remain unanswered questions. We have been conducting a series of studies examining variables which might affect this source of data. Again

major work has been done by the team of McGill graduate students which included Lynn McAlpine, Vivian Medley-Marks, and Charles Burt.

Burt's doctoral dissertation (Burt, 1988) examined the effects of different roles of the student test subject and the tester on the *amount* and *kinds* of data obtained from the developmental testing of instructional materials. He also looked at varying the *number* of participants in the session and at the accumulation of *new data* as a function of different data gathering procedures. He was stimulated to do this research because of the quite different suggestions in the literature concerning what the testing procedures should be. Some authors see the activity as a simulation of the student using the final version of the materials and so they suggest that students simply work through the learning materials and take a test afterwards. The gain between pre- and post-tests will indicate the degree of effectiveness, and the particular errors on the post-test will provide diagnostic leads. Another group of authors suggest a more clinical approach, and view the test subject as a partner in the revision process. They propose a much more informal, interactive, and personal set of procedures.

Sessions in Burt's study involved one or several test subjects. The role of the the subject varied from *passive* (taking a pre-test, going through the materials while making comments, and taking a post-test) to *active* (engaging in dialogue with the tester during the test session). Similarly the tester could be *passive* (merely recording events) or, *active* (probing for comments and explanations). The information gathering procedures included testing before and after the session, attitude questionnaires, and debriefing after the session. One important finding was that a very few subjects contributed most of the commentary data. Most test subjects were silent Sams and Samanthas. This suggests that some screening should precede selecting the subjects and/or that the preparation for the task should include some sort of training or practice. Just as experts differ, so students are quite unlike as data generators. A second and

have utilized think aloud techniques in diagnosing learning difficulties. Like mine, your grade school teacher probably used to caution: "Show your work as well as your answer."

somewhat surprising finding was that each additional probe for data (e.g., the test, the attitude surveys, the debriefings) produced a good deal of new information. The various techniques of gathering data were not redundant.

Obviously there is more to describe about each of these two dissertations and there are several other relevant short studies that should be mentioned as well. But the purpose of this paper is not to do that. It was really meant to encourage you to say four things. I will end by noting them.

1) There is evolving a systematic technology of instruction. Though clearly based on psychological principles, it is not merely an application of them to specific problems of teaching. The technology deals with *how* such applications can be made using replicable methods and, though expressed in various educational innovations, it is not specific to any particular innovation.

2) The technology differs importantly from traditional instructional development not only because it espouses certain principles of learning but also because it *is* systematic; it is driven by a commonly agreed upon, explicit heuristic.

3) Of the components of that heuristic one in particular distinguishes it from

typical instructional development (and makes this technology similar to other engineering efforts). The iterative cycle of testing provides feedback for improvement.

4) There is research being carried out, within and outside of universities which directly addresses generic concerns raised by the technology.

Obviously I agree, Fred, that there is a *fifth* thing that we should continue to say: "I don't know."

REFERENCES

- Burt, C. (1988). *Identification of critical variables in developmental testing and an experimental examination of the number and roles of participants in testing sessions*. Unpublished doctoral dissertation, McGill University, Montreal.
- Geis, G. L. (1986). Formative evaluation: The receiving side. *Performance and Instruction*, 25(5), 3-7.
- Geis, G. L. (1987). Formative evaluation: Developmental testing and expert review. *Performance and Instruction*, 26(4), 1-8.
- Markle, S. (1978). *Designs for instructional designers*. Champaign, IL: Stipes Publishing Company.
- Saroyan, A. (1989). *The review process of formative evaluation: The role of the content expert and the instructional designer*. Unpublished doctoral dissertation, McGill University, Montreal.
- Saroyan, A., & Geis, G. L. (1988). An analysis of guidelines for expert reviewers. *Instructional Science*, 17, 101-128.