A Case Study: Developing Convergent Formative Evaluation Methodology

The following case study has been severely edited due to space constraints. An effort was made to maintain the flavor of the case study, to ensure the results are presented and to mention the statistical procedures used. A complete report of the project is available from the authors. *JBJ*

Thomas M. Schwen
Director
Division of Development and Special Projects
Indiana University

John M. Keller
Assistant Professor of Education
Syracuse University

This project was funded by the Division of Development and Special Projects, Audio Visual Center, Indiana University

Introduction

This is a case study of the evaluation of a course development project. Our goal was to develop an evaluation design which would provide convergent data in making decisions about revisions in the course materials and exercises. Our initial problem was that no *a priori* criteria existed to use a basis for a criterion referenced evaluation, nor was it possible to use any comparison groups.

To put the evaluation in perspective, we will present some background information about the human geography course, the early course development work and the instrument development process. Then we will present the design and results of the evaluation after which we will provide a general discussion of the outcomes.

Course Description Human Geography is an introductory course which typically has an enrollment of 150 studentsd, mostly freshmen and sophomores. The first eleven weeks were structured in the typical auto-tutorial manner: a one hour large-group lecture at the beginning of the week; a one to two hour lesson in the auto-tutorial lab; and a one-hour small-group discussion section at the end of the week. The remaining three weeks of the semester the students developed a term paper from raw data and journal articles supplied by the instructor. The students usually attended discussion sections for assistance with the paper. Also, some students reviewed previous lessons. All lecture sessions were cancelled during that three week period.

During the first eleven weeks a typical lesson would be introduced at the large group meeting on Monday with either introductory comments or a film. At the learning lab (the auto-tutorial study center) the lesson was largely self-contained, however, an Associate Instructor (A.I.) was in attendance and performed specific teaching and evaluation functions. The discussion session at the end of the week involved a planned activity. These sessions, which were led by the A.I.s, included games, simulations, and semi-programmed activities designed to extend the concepts of the lesson into an applied or problem solving situation. These discussion sessions were designed to aid in the transfer of the concepts, principles, and skills presented in each auto-tutorial lesson.

In the final three weeks, the concern for transfer of course learning was extended in the Applications Phase of the course. The student was expected to utilize the packet of unfamiliar raw data and journal articles in a manner that demonstrated knowledge of course objectives. The instructions to the students and criteria for judging the paper required specific "applications" of the concepts, principles and skills in treating the new content. We often used the phrase, "the students must think like a geographer," to describe this activity.

Products and Equipment

Large Group Lecture Session The first session was concerned with the usual matters of orientation. The sessions consisted of films, lectures, and testing as follows: (1) five used films which had been selected following an extensive bibliographic search, (2) five were of a traditional lecture format, and (3) three were devoted to testing. There were two weeks in the semester during which no lecture session was held.

Auto-Tutorial Laboratory Each of the eleven weekly sessions consisted of a printed student guide, a tape recorded lesson, and, in all but one case, a filmstrip. Other visual materials such as maps, charts, and graphs were included in each mimeographed student guide. Each student guide had four basic elements: (1) an advanced organizer, (2) behavioral objectives, (3) student worksheets, and (4) selected readings. The tape guided the student through a semi-programmed set of experiences. When played without repetition, the tape ran from 30 to 45 minutes. The filmstrips were developed to illustrate concepts, or to provide stimuli for skill-guiding exercises. The filmstrips were prepared from original slides supplied by the instructor.

Discussion Session The A.I.s led the sessions following a reasonably precise lesson plan. Three of the sessions utilized simulations and games adapted from the High School Geography Project. Three of the remaining sessions were devoted to test preparation. The remaining sessions used a guided inquiry technique in the solution of assigned problems.

Applications Phase The Applications Packet included a newsletter, which listed the objectives for the final paper and the criteria and process by which the paper would be judged. The data in the packet were of two types: (1) census and other demographic information,

and (2) a few journal articles, which contained information that provided a basis for application of the analytic concepts and skills learned during the auto-tutorial lessons.

The Developmental Analysis

During the first four to five months of the project, the behavioral objectives, the test items and the task analyses were revised. The instructor was particularly concerned with assessing "higher order" cognitive outcomes and using the data to improve his instructional materials. Fifty-five behavioral objectives were developed that specified three types of behavior: (1) higher order concepts, (2) five groupings of principles termed organizers, and (3) four skills. All concepts were complex or of a "higher order" in that each definition included at least two subconcepts. The groupings of principles were simple descriptive models (Chorley & Haggert, 1967). The models described the relationships between four or more concepts. The four skills involved three types of mapping and the interpretation of scattergrams.

Development of Achievement Tests The first achievement tests were developed in the Summer of 1973. Three, twenty-five item tests were developed in the tradition of criterion referenced measurement. Test validity was largely a logical process (content validity) in which test items were matched to behavioral objectives. We also analyzed correlational data utilizing aptitude, and grade point average and time in the laboratory in the determination of validity (see discussion below).

The distributions of achievement test scores generally conformed to standards for criterion referenced tests. Considering all forms of course performance, 75% of the students exceeded 80% of the criterion score. Considering the three criterion tests individually, in two of the tests 70% of the students exceeded 80% of the criterion. Early in the process of developing the third test only 45% of the students exceeded 80% of the criterion. The objectives, the lessons, and the test items were modified for this section of the course. In subsequent replications the data were more satisfying, 75% of the students exceeded 80% of the criterion score.

Evaluation Design and Results

Design It was the general goal of the evaluation to develop convergent feedback that would lead to appropriate revisions in the development of the materials and exercises. This general goal was restated in terms of three questions.

The first question: Do student perceptions regarding the first eleven weeks of the course lead to meaningful corrections in the materials and exercises?

The second question: What student individual differences and task related behaviors are correlated with success in the course? This question was posed with two general consequences in mind. Our attempt to correlate individual difference measures with success would allow us to examine the expected pattern of academic aptitudes being positively correlated with achievement. Our interest in task related behaviors such as repetition in the lesson, reading objectives, etc., allowed us to empirically examine the effectiveness of these tactics. We felt that empirical confirmation of these data would allow us to be more confident in our advise to students and in cross checking our other sources of information. Finally, the sources of information (individual differences and task related behaviors) used in the same analysis allowed us to assess the effects of individual differences versus instructional tactics. We expected the variance accounted for to be somewhat independent in a criterion referenced learning situation. Individual differences in academic aptitude will invariably predict success to some extent. However, we anticipated that the use of our instructional strategies should account for independent variance in achievement. In other words, the strength of the treatments would, partially, overcome the effects of aptitude or previous learning.

The third question: Can the students articulate the relationship between the first eleven weeks and the Applications phase of the course? A major portion of our effort was directed to this question because the question reflects the most important objective of the course. Our rationale was, if the students were to "think like geographers" then, they must understand or be able to articulate the relationship between the two major phases of the course.

Since each of the questions involved a primary methodology and the resulting activities were independent of the others, they will be treated separately in the following descriptions. However, we were able to examine the validity of our observations by comparing the results from each type of instrumentation and by comparing these results with other data such as the intuitive judgments of the instructor. We came to regard the intuitive judgments of the instructor as a useful, semi-formal source of information. As the case proceeded we would seek his opinion while minimizing information that would bias his response. We subsequently felt this procedure sharpened our evaluation process. The process of creating overlap among the types of instrumentation was viewed as an informal extension of the concept of convergent validation (Campbell and Fiske, 1959).

Question 1: Lesson by Lesson As previously discussed, the Introductory Phase of the course included eleven weekly lessons and each lesson consisted of a large group lecture, an auto-tutorial lesson, and a small group discussion section. The weekly evaluations were concerned with the auto-tutorial portion of each lesson. During the fall semester, 1973, the first six lessons, and lessons 10 and 11 were evaluated by each student responding to a questionnaire immediately following each lesson. The questionnaires included 10 Likert-type items with five response choices ranging from strongly agree to strongly disagree. Items were worded in a manner designed to counteract positive or negative response bias.

Each weekly questionnaire contained ten questions on a variety of topics. Some topics were included in each questionnaire, while others were included in only one or two of the weekly evaluations. The complete list of topics follows:

1. Repeated portions of the tape: This category sought to determine whether the students did replay portions of the tape recording used in each lesson.

2. Student guide: The student guide includes instructions to the student, the objectives of the lesson, brief skill-building "mini-lessons," readings, and exercises. Questions in this category concerned evaluations of the student guide.

3. Orientation value of introductory outline: Beginning with Lesson five, an introductory outline was added to the student guide. Students were asked their opinions on the usefulness of this outline as an orienting device.

4. Relationship between performance objectives and content: Students were asked whether the performance objectives were clearly stated, and whether the content of each lesson was clearly related to the performance objectives.

5. Student self-assessment: Students were asked to indicate the degree to which they felt they had understood and accomplished the objectives of the lesson.

6. Tape/voice quality: Since poor tape quality can interfere with learning, this characteristic was evaluated independently of the tape content.

7. Auto-tutorial instructor: At some point during a lesson the student was asked to check, or discuss, his results with the learning center instructor. This category evaluated the students' attitudes toward this part of the instruction.

8. Instructional techniques: A variety of instructional techniques were used, and the students were asked to evaluate their effectiveness.

The responses to each item in each category were presented in tables by percentages. Some tests of statistical significance would not provide a strictly objective, probabilistic basis for decision making. We decided to formulate a set of arbitrary decision rules to use as a guide to interpreting what actions to take on the basis of the results. We decided that if the total percentage of responses to the Agree and Strongly Agree responses fell below 50% for a given item, two actions were taken: (1) the item was re-examined for possible sources of ambiguity, and (2) if the item appeared to be reliable, further inquiry into the problem represented by the item was recommended. If the average agreement to all items in a category fell below 50%, then further inquiry into problems associated with that category was recommended. Total agreement between 50% and 70% was considered to be satisfactory support, and total agreement above 70% was considered to be strong support for an item.

Question 2: Achievement Correlates
The purpose of this phase of the evaluation was to identify individual differences in the students' behavior that were correlated with success in the course. Fourteen behaviors were identified as possibly contributing to differences in achievement measures. The relationships between these variables and test scores were analyzed while controlling for several aptitude and status variables.

In this phase, 36 of the 156 students in the course were used in the evaluation. Twelve students were randomly selected from each of the high, medium, and low scoring thirds on the first test. They were interviewed by telephone during the fifth week of the course. Data were also obtained from class records at the conclusion of the course. The data from the interview were analyzed by means of crosstabulation with test score group using the Crosstabs program in Statistical Packages for the Social Sciences (Nie, Bent, & Hull, 1970). The data obtained from class records were of a nominal order, therefore, product-moment correlations and, in a few cases, multiple-regressions were used for the analysis, again using programs from SPSS (Nie, *et al*, 1970).

Question 3: Follow-up Evaluation The client wanted to know, independently of the actual results of the assignment, whether the students felt that they understood the applications assignment and its purpose. Another closely related aspect of this question was whether the students perceived and understood what the relationship was betwen the first eleven weeks and Application Phase of the course. A third aspect of this question was with the students' general reaction to the auto-tutorial method of teaching the lessons. An effort was made to determine the extent to which the students regarded it as having been effective, efficient, and interesting. Of course, the interview information overlapped the other sources.

Results

To present the results in their entirity would be lengthy, tedious, and beyond the primary purpose of this report. It is our purpose to present some of the more interesting findings in a manner which illustrates the design of each phase of the evaluation, and which demonstrates the potential contribution which these procedures can make to a development project.

Question 1: Lesson-by-Lesson The results of the first six lessons indicated that the content of the lessons was well-organized, the objectives were clear, the instructional material was clearly related to the objectives, and the instructional techniques were effective. However, the results from Lessons 10 and 11 were not so positive. In these lessons a substantial percentage (50%) of the students indicated that they did not have a complete understanding of some of the concepts, and that the lesson content was not always clearly related to the objectives. This feedback was interesting for several reasons. First, it isolated a source of confusion in the materials from the students' point of view (they performed satisfactorily on the achievement test, but we were also concerned with the students' perception of the lesson objectives, and that they perceived that the lesson strategies were consistent with the objectives). The second point of interest is that these results confirmed the intuitive reaction of the instructor with regard to these lessons. He had felt that these difficulties still existed, and that further revisions were necessary. The final point of interest is that these results tended to validate the successful reactions to the first six lessons. We concluded that the apparent honesty of the students in critically evaluating their reaction to the materials in this case supported our assumption that they were being honest in their more positive responses to the earlier lessons. The client indicated that the confirmation of his perceptions which these evaluations provided gave him a firmer grasp of the lesson design techniques which led to successful lessons.

These weekly evaluations also allowed us to identify an additional area of weakness in the course. One function which the A.I.s were expected to perform in the learning lab was to give the students feedback and assistance with specific exercises in the auto-tutorial lessons. The students were instructed at specific points in the lesson to show their responses to an exercise to the A.I.s and discuss their results with them. The responses to this item when it was included in the weekly evaluations indicated a balance of agreement with disagreement, with fully a third of the respondents undecided as to whether or not the learning center instructor was helpful. With this feedback it was possible to conduct further inquiry aimed at the specific problem of determining whether (1) the students felt that they did not need any help, or (2) the learning center instructors were either unclear as to their role, unprepared, or unwilling to provide the appropriate, helpful assistance which was desired.

Question 2: Achievement Correlates
The most significant finding from this portion of the evaluation was that almost all of the high and medium scorers on the first test read the objectives for each of the first three lessons before turning on the tape, while less than half of the low scorers did so. Most of the low scorers read the objectives only occasionally. The higher scorers were also slightly more likely than the low scorers to read the outline of the lesson before beginning, to discuss their work with the A.I. in the lab, and do advance preparation for the discussion group meeting. There were no noticeable differences among score groups on any of the remaining variables included in the interview. One reason for this was simply that there was no variance on several of the items. Over 85% of the subjects, regardless of test score, indicated that they (1) took notes in the student guide, (2) performed the activities indicated in the

lesson, (3) repeated portions of the tape, (4) did not skip portions of the tape, (5) completed each lesson before the discussion group met, (6) attended all the discussion groups, and (7) participated in the discussion groups. Also, 90% of the students attended the lab when a G110 A.I. was on duty. There was a variance as to whether or not the students perceived that the discussion group provided practice of the lesson's objective (70% said yes, and 30% no), but the differences between score groups are difficult to interpret. There was actually a greater tendency for lower scorers to have responded positively. The greater tendency to respond negatively on the part of the higher scorers could be a result of their being more critical as a result of greater study and greater familiarity with the objectives.

In the remaining portion of this phase of the evaluation, several objective measures of the students' behaviors were taken from class records and the university data file, and compared with achievement in the class. This was done during the fifth week—one and one-half weeks after the first test, and at the end of the semester after final grades had been posted.

The first part of this analysis was concerned with the relationship between attendance and participation in discussion groups and total test score. Points were assigned to each student by the A.I.'s for attendance and participation. It was found that attendance was not an effective predictor because almost everyone got the maximum number of points (the maximum possible was 13, and the mean was 12). On the other hand, participation was significantly ($p < .10$) correlated with total test score ($r = .24$). When SAT scores were partialled out of the correlation, the coefficient increased slightly ($r = .27$). This means that participation in the discussion groups is positively related to test score independently of ability as measured by SAT.

The major part of this analysis was concerned with the relationships between time-in-lab, SAT, GPA, and test score. These relationships were studied after the first test, and at the end of the semester. Scores on the first test were found to correlate significantly ($p < .10$) with time-in-lab, SAT, and GPA. The correlation between time-in-lab and test score increased from .35 to .40 when the correlation due to SAT was partialled out of the equation, but it decreased to .22 when GPA was partialled out. This suggested that while time-in-lab and SAT are both correlated with achievement, their relationships are independent; that is, effort expended on the developed materials appears to have a correlation with achievement which is independent of, and greater than ability as measured by SAT. However, time-in-lab is not independent of GPA, which itself is probably more of an indication of effort than ability.

At the end of the semester, these relationships did not hold up. Although all three independent variables were still significantly correlated with test score, the time-in-lab correlation had shrunk while the other two increased. Time-in-lab was still significantly correlated with test score after SAT had been partialled out, but there was no correlation when GPA was partialled out. Therefore, time-in-lab appears to be independent of ability, as measured by SAT but not independent of other aptitudes, personality characteristics, or efforts which have been employed by the higher achievers in other courses.

Question 3: Follow-up Interview The results of this interview, obtained six weeks after the end of the semester, indicated that over 80% of the students recalled what the final assignment was, and what its purpose was. Only 58% of the students sought the aid of the A.I.'s while working on the final paper, but all of those who did seek help said that the A.I.'s were helpful. The positiveness of this response was an improvement over the Introductory Phase when the students often perceived that the A.I.'s were not helpful. It may be that the A.I.'s had a more useful function to serve during the Applications Phase.

Over 90% of the students recalled that there was an instructional relationship between the Introductory and Applications Phases of the course, and they were able to describe what this relationship was. And, over two-thirds of the students expressed a favorable attitude toward the way in which the course was organized.

Again, over 90% of the students expressed a favorable attitude toward the individual, auto-tutorial lab as a means of teaching the lessons. An even greater percentage (97%) indicated that the auto-tutorial lab provided an effective and efficient means of learning the material. However, when asked if working in the lab was enjoyable, 47% said yes, while 43% were indifferent, and 10% said no. Even so, over 90% said that they would not have preferred another method of presenting the lessons.

Discussion

We interpreted these results as indicating both strong positive support for (1) the auto-tutorial instructional technique, (2) the other instructional techniques used in the course, and (3) the overall instructional quality of the materials which had been developed for the course. We were impressed by the ability of the randomly selected students who were interviewed in the Follow-Up evaluation to recall specific goals and objectives of the course, as well as its general structure.

With respect to the primary goal of the evaluation, we were able to identify some specific problem areas in the course which aided in the development, and we were able to identify several correlates of successful achievement, which were related to effort within the course. One example of a problem area consisted of the student evaluation of Lessons 10 and 11. The weekly evaluations of Lessons 10 and 11 indicated that the students saw a lack of relationship between some of the exercises and the objectives of the lesson, and that some of the concepts were not clear to them. There were two possible explanations for these results. One was that the students were correct, and the other was that they were being highly critical due to having had considerably lower grades on Test 2 than Test 1. We were able to make a decision by considering two additional sources of information. It may be recalled that time-in-lab, which presumably is a measure of the amount of time spent studying the materials lost some of its predictive power betwen the first and third tests. This could indicate that during Lessons 10 and 11, the students had to fall back on learning strategies which they had acquired prior to entering the course. The other source of information was the instructor. When these results were presented to him, he indicated that he had suspected that there might still be difficulties in these lessons. He then undertook revisions on the basis of the evaluative data.

Achievement in the course was found to be positively correlated to several behaviors which were measures of effort within the course. These included studying the objectives before beginning the lesson, participating (rather than just attending) the discussion groups, and amount of time spent in the auto-tutorial lab. The correlation of amount of time spent in the lab and achievement was in-

dependent of SAT throughout the course, and partially independent of GPA on the first test. These relationships were reported to subsequent students who were enrolled.

Our primary goal seems to have been met. Our information seemed to converge on meaningful corrections in materials or exercises. As has already been mentioned, the methodology provided the instructor with a means of validating his perceptions of the outcomes of the materials he had developed, in combination with multiple sources of validation of several findings. An additional outcome is that the design used here is proving to be generalizable to other course development projects. The combination of (1) weekly student evaluations, (2) midterm telephone interviews with a random sample of students, (3) a follow-up telephone interview with a random selection of students, (4) achievement data, and (5) careful use of intuitive observations of the instructor and associate instructors is not expensive and does not require elaborate or sophisticated technology to implement. It does provide a basis for multiple, or convergent, methods of validating given outcomes, and for making comparisons of outcomes over time as the course is repeated in successive semesters.

References

Campbell, D.T., & Fiske, D.W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 1959, 56, 81-105.

Campbell, D. T. & Stanley, J. C. *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally, 1963.

Chorley, R. J. & Haggert, P. *Models in geography*. London: Methuen & Co., Ltd., 1967.

Nie, N. H., Bent, D. H., & Hull, C. H. *Statistical packages for the social sciences*. New York: McGraw-Hill, 1970.