Problems in the evaluation of PLATO lessons

ELSBETH M. HOLT
Centreville, Virginia 22020

The evaluation of the effectiveness of PLATO CAI in both educational and industrial settings is discussed. The evaluation of the PLATO instruction in the educational setting generally supports the effectiveness of PLATO instruction but was hampered by limited control of instructional method for different populations of students. Further difficulties in evaluation were caused by ceiling effects in the measurement instrument. Evaluation of PLATO instruction in the consulting setting was hampered by lack of cooperation by the client institution. Mastery of the content of some areas covered by the PLATO lessons in this setting was high, whereas mastery of other content areas was low. Conclusions address the necessity for at least a minimal amount of cooperation from the client institution and a minimal amount of control of the evaluation setting for adequate evaluation of any CAI instructional method.

In this paper, I will discuss some of the problems encountered in trying to evaluate PLATO IV computer-based education (CBE). Theoretically, all material should be evaluated to find out whether it teaches what we want it to teach. However, in academic and in applied settings, there are evaluation problems.

I have been involved in designing, programming, and evaluating PLATO IV instruction both in academic and in industrial settings. Some of the problems and successes we have had in trying to run a theoretically sound evaluation on the effectiveness of our material in both settings are discussed here.

EDUCATIONAL SETTING

At the College of Veterinary Medicine at the University of Illinois, we developed a PLATO program that became more and more an integral part of the instruction. Some of the material supplanted lectures, other material was used in conjunction with traditional instruction, and some was to be used as enrichment for the students.

In the veterinary public health area, the complete course was ultimately converted to CBE. In the process,

it was necessary to design routing and testing programs. Some of the instructional material consisted of tapes; other material was PLATO computer-assisted instruction (CAI). However, before we did the conversion, some of the CAI lessons were tested. In the first study (done in the spring semester, 1973), 18 students were matched on grade-point average (GPA) to form nine pairs of students (correlation between GPAs in the two groups was .99). The standard academic lecture that the classroom group attended contained exactly the same material as the PLATO lesson. Five students did not come to class and thus received no instruction, a self-selected control group.

The column labeled "1973" in Table 1 shows the results of comparing the different groups. Some instruction was better than no instruction, but PLATO instruction was significantly better than the matched or the total classroom group.

In 1974, we wanted to test the effectiveness of the two CAI lessons in veterinary public health. We had planned to repeat the test from the previous year for both lessons, but we ran into a serious problem. The students were antagonistic to the idea of being randomly assigned to a group. Part of this was probably due to

Table 1
Average Grades on Tests in Public Health

	1973		1974a		1974b		Comparison	Results		
	N	Mean	N	Mean	N	Mean		1973	1974a	1974b
(1) PLATO Only	9	74.44	20	91.82	34	97.06	(1),(2)	t(8) = 2.56*		
(2) Classroom (Matched)	9	53.33					(1),(3)	t(42) = 2.98**		
(3) Classroom (Total)	35	52.31	27	86.75			(1,2),(6)	F(2,51) = 9.9**		
(4) PLATO and Classroom			29	93.28			(1,4),(3)		F(2,73) = 3.95*	
(5) Handout					43	96.08	(1),(4)		Nonsignificant†	
(6) No Instruction	5	27.00					(1),(5)			Nonsignificant†

* $p < .05$. ** $p < .01$. † $p > .05$.

some very basic antagonism toward the PLATO system (these students had lived through some of the teething problems of the PLATO IV system in 1973); the other part was a worry that one group might have an advantage over the other. Veterinary medicine students are very competitive as a group and do not want to be randomly assigned to a possibly inferior mode of instruction. The problem was solved by letting the students choose the type of instruction. They could (1) attend classroom lectures only, (2) study the PLATO lessons only, or (3) do both for Part A of our study. For Part B, we transferred the text material of the PLATO lesson to a handout. Each student could decide whether to study the handout or the PLATO lesson. The results of these studies are shown in Table 1, in the columns labeled "1974(A)" and "1975(B)."

For Part A, a significant difference in average grades was found between the PLATO-only and the classroom groups, and between the PLATO-and-classroom group and the classroom-only group. There was no statistically significant difference in grade between the PLATO-only and the classroom-and-PLATO group. The GPAs of the three groups were compared to determine whether there was any statistically significant difference between the three groups: None was found.

The second study of 1974 pointed out another problem in evaluating CAI vs. other instruction: There was a ceiling effect. Both groups had mastered the material to such a high degree that no statistically significant difference between the groups could be established. We had to be satisfied with the knowledge that all of our students mastered the material.

Let us look at two more studies that were done in the fall of 1974. A PLATO lesson "Histology of the Skin" was developed as a study program to cover the laboratory portion of the instruction on this topic. In the fall semester of 1974, the freshmen taking this course were divided into two groups: students who studied the PLATO program, and a second group that attended the traditional laboratory at the same time. The instructor felt that restricting the use of PLATO would not be fair; thus the students of the second group were free to voluntarily study the PLATO lesson. Of the laboratory students, 69% chose to review the material on PLATO, reflecting the great motivation and competitiveness of veterinary medicine students. The test told the same story. Over 60% of all students missed at most 2 of 30 questions (93.3% or higher). With such a ceiling effect, differences in the groups were hard to detect, and we were not surprised to find no statistically significant difference between groups (see Table 2). All students, whether they used PLATO only, went to the laboratory only, or studied both ways, reached an extremely high level of mastery of the material.

The last study to be mentioned here took place during the summer and fall of 1974. Each incoming

Table 2
Average Grades on Tests in Histology and Terminology

	Histology*			Terminology†		
	N	Mean	SD	N	Mean	SD
PLATO Only	43	93.3	5.60	40	77.1	12.5
LAB Only	13	90.8	9.04	47	71.3	12.3
PLATO and LAB	29	92.2	6.40			

*Results were nonsignificant. † $t(40) = 2.1743$, $p < .05$.

freshman was sent a letter with information about PLATO. The students were invited to work through two programs on veterinary and anatomical terminology as practice in basic veterinary language and phraseology. The first quiz of the semester tested the student's knowledge of definitions, both lecture and reading material, without being directly related to the PLATO lessons. It was thought that studying the terminology programs might help the student in reading and interpreting the lecture and reading material. Forty of 87 freshmen (46%) used the lessons voluntarily before the first quiz. The mean scores for this group were significantly higher than the scores for the rest of the class (Table 2). Before we credited all of this to the effectiveness of the PLATO lessons, we waited until the end of the semester and compared the GPAs of the two groups. The average GPA of the students that had voluntarily studied the PLATO lessons was significantly higher than the average GPA of the rest of the freshman class. Thus there is a very real explanation that the higher scores on the first test are due to having a self-selected sample of students with higher motivation to study. Another possibility is that the two PLATO lessons helped the students in all of their classes during the first semester.

In summary, I would suggest that PLATO instruction can be as effective as classroom instruction. The problems encountered in devising sound testing of effectiveness of the PLATO lessons can be traced to the highly competitive atmosphere at the College of Veterinary Medicine: Students are unhappy with being randomly assigned to groups; they would rather have all possible instruction, or decide themselves how they want to study. Further, the students always strive for complete mastery of the material no matter how it is presented; thus ceiling effects make the discovery of true differences very difficult.

INDUSTRIAL SETTING

For the last 2 years, I have worked in an industrial setting. We contracted to convert four existing courses that had overlapping material and were taught using traditional instructional methods into one computer-managed instruction (CMI) course. The plan was to use the CMI package on the PLATO system to route the

student to the appropriate learning resources and tests. The learning resources consisted of CAI lessons, slide tape presentations, videotapes, and textual material. All testing and routing was to be done using the PLATO CMI system and criterion-referenced testing to establish mastery of objectives.

The original contract specified a formative evaluation model to be used to evaluate the material and test items. The model involved four steps: (1) expert review, (2) one-to-one testing, (3) pilot testing, and (4) field testing. While our company developed the material, the client was supposed to furnish the people to carry out the four steps in the formative evaluation. They did have experts review the material, so that we have some confidence that we have some face validity. However, we were never able to find a student with sufficient time for the completion of the one-to-one testing. We did, however, run a pilot test on the first two modules that were developed as a preliminary product to show our capabilities.

We had hoped to be able to use 25 students who had similar characteristics as the people who would later receive the instruction. Instead, only 12 people were provided, and their characteristics varied widely from the people who would have to take the course in the future. For instance, a typical student averaged 3.17 years in the client organization; the pilot test participants averaged 10.15 years.

We collected data on mastery of the modules for each participant (Table 3) and allowed a maximum of three tests for each module. For Module A, 90% of the students mastered the material; only 60% of the same participants mastered Module B in three or fewer trials. There were several possible explanations for this result: (1) The lessons were ineffective; (2) the tests did not really test mastery; (3) the students did not have enough background for the material covered in Module B, whereas the material in Module A depended less on prerequisite knowledge. To distinguish between the first two possibilities, we collected response distributions to the items on the first test for all subjects. Table 4 shows the distribution for the questions for one of the objectives for Module B. It shows that this type of data

Table 3
Distribution of Student Mastery for the Two Modules of the Pilot Test

Number of Students	Module A (N = 10)	Module B (N = 10)
Mastering First Test	6	2
Mastering Second Test	1	4
Mastering Third Test	2	0
Not Mastering After Three Tests	1	4

Table 4
Distribution of Answers for the Questions for One of the Objectives for Module B on the Pilot Test

An- swer	Question (N = 10)										
	1	2	3	4	5	6	7	8	9	10	11
a	5*	6*	3		1			3	3	2	1
b	5		1	5*	2	8*	3*	4*	3	4*	4
c		3	3	4		1	2		1	3	
d		1	3*	1	7*	1	5	3	3*	1	5*

*Correct answer for that question.

does give some idea concerning which questions were major causes for nonmastery. However, it still does not tell whether the material was not taught correctly or whether the question was ambiguous. Expert review of the questions and instructional material is needed.

The pilot test thus provided some insight into the evaluation of our instructional material. However, since there were only 10 participants, the results were not generalizable.

The final point in the formative evaluation model is the field test. At one time, our client was committed to run a prototype course, gathering as much data on performance and results as possible using true students as participants. This has not been done because of budget restraints. So it is extremely difficult to determine the effectiveness of our training material.

In summary, good evaluation is possible only if the investigator has control over the student or subject population and if that population is willing to cooperate in the specified assessment design.