DOCUMENT RESUME

ED 204 403 TM 810 417

AUTHOR Upp, Caroline M.; Barcikowski, Robert S.
TITLE The Evaluation of a Model for the Assessment of Class Progress.
INSTITUTION Ohio Univ., Athens.
SPONS AGENCY Ohio State Dept. of Education, Columbus.
PUB DATE Apr 81
NOTE 51p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS *Academic Achievement; Attitude Measures; Cognitive Measurement; *Evaluation Methods; Formative Evaluation; Higher Education; Instructional Improvement; *Item Sampling; *Models; *Program Evaluation

ABSTRACT

Demands for more complete information on educational programs have emanated from national, state and local sources. Their focus is on the processes that are occurring in individual classrooms. The information that is collected to provide insight into educational programs is customarily summative in nature, answering, for example, questions regarding student progress toward accomplishment of objectives. Here, a model is evaluated which allows that decisions be made for the benefit of the students still enrolled in classes, not merely the students of future classes. This model, the B-model, is concerned with cognitive and affective outcomes. Results show that this model provides researchers and school administrators with a sensitive measurement approach, which is economical in terms of teacher/student time, with which to measure student progress and, perhaps, teacher effectiveness throughout the school year. (Author/GK)

THE EVALUATION OF A MODEL FOR THE ASSESSMENT OF CLASS PROGRESS[1]

Caroline M. Upp and Robert S. Barcikowski
Ohio University

Paper presented at the annual meeting of the
American Educational Research Association,
Los Angeles, April, 1981

[1] The preparation of this paper was partially supported through an Ohio State Department of Education Project 419 - Year II Research and Development Grant. The opinions expressed herein do not necessarily reflect the position or policy of the Ohio State Department of Education, and no official endorsement should be inferred.

ABSTRACT

Demands for more complete information on educational programs have emanated from national, state and local sources. Their focus, in the final analysis, is on the processes that are occurring in individual classrooms. The information that is collected to provide insight into educational programs is customarily summative in nature, answering such questions as "What is the average reading level of fourth grade students in May?" Here, a model is evaluated that requires measurements throughout a school term so that decisions can be made that will benefit the students still enrolled in classes, not merely the students of future classes.

The Evaluation of a Model for the Assessment of Class Progress

Introduction

Evaluation in the field of education has been defined by Cronbach as "the collection and use of information to make decisions about educational programs" (Cronbach, 1963). Such evaluation has been an aim of educators and laymen alike. At local, state, and national levels, demands for information about the effectiveness of educational programs are heard. Evidence of these demands can be found in Michigan's state-mandated accountability program (Porter, 1976), the ninth annual Gallup Poll of public attitudes toward education (Gallup, 1977), and concern over declining test scores (Ebel, 1976). Legislative bodies in state houses and the Congress have authorized funds to be used expressly for evaluating educational programs (Worthen & Sanders, 1973).

These demands for more complete information on educational programs have emanated from national, state, and local sources. Their focus, in the final analysis, is on the processes that are occurring in individual classrooms. The information that is collected to provide insight into educational programs is customarily summative in nature, answering such questions as "What is the average reading level of fourth grade students in May?" or "How well did students taking this year's ACT tests perform as compared with those students who were tested in the same manner five or ten years ago?"

These questions are important ones, to be sure. However, there are other questions that are of at least equal importance to the individual classroom teacher and to the local school system. These are such questions as the following:

1. How is this particular group of students progressing toward accomplishment of the objectives of the term?
2. As the school term proceeds and new topics are introduced, are the students retaining their learning from previous weeks?
3. Has a point been reached at which the curve of accumulated learning is flattening or descending?
4. What are the attitudes of the students toward the subject at hand?
5. Are these attitudes changing? If so, are they becoming more positive or more negative?
6. Is this group of students learning at a rate comparable to that of similar groups?

These questions cannot be answered by summative measurements taken only at the end of the term, but rather must be answered by means of frequent testing throughout the school term. If such measurements are taken throughout the term, decisions can be made that will benefit the students still enrolled in the class, not merely the students who will be enrolled in future terms.

The principal barriers preventing the collection of frequent measurements have been concerned with the omnipresent factors of time and money. Frequent testing complete enough to provide accurate information on students' cumulative progress toward yearly goals has been extremely costly in terms of teacher and pupil time and testing expense. Barcikowski and Upp (1978), however, have suggested an approach, referred to here as the B-model, based on multiple matrix sampling which may enable frequent, accurate collection of such data at a fraction of the customary cost. The utilization of a multiple matrix sampling process requires that only a few test items need be administered to each student. Because the questions to be answered by the testing program refer to group progress, and not to individual achievement, accurate estimates can be derived from this small number of items administered to each student. For example, instead of a test of 150 items for each student, a test of twelve to fifteen items per student may be sufficient to provide a reasonably accurate measurement of the achievement of the class. Computer compilation, printing, and scoring of the tests provide accurate data on class status with minimal input of teacher time. The B-model is designed to measure progress toward both cognitive and affective objectives for the term in this fashion.

Class Progress: What Should be Measured?

Before a measurements system can be initiated for evaluating class progress, some decisions must be made regarding the aspects to be evaluated. The outcomes of education are multidimensional, as is the process of education itself. Some of these dimensions are cognitive; some are affective; others relate to moral character, adjustment to life, self-confidence, and citizenship. In evaluating the performance of a class of pupils, any of these dimensions may be assessed. The general public demands assessment of cognitive outcomes (Porter, 1976; Gallup, 1977; Ebel, 1973); and it does appear that any evaluation of class progress must give attention to cognitive skills.

One of the goals of teaching, however, is to encourage students to go beyond the subject matter being covered in class and to encourage deeper and wider pursuit of the subject. A study by French (1961) indicates that favorable student attitudes do lead to increased time spent at that activity and choice of further scholastic pursuits related to that field. It therefore is desirable to have favorable student attitudes toward the subject matter being studied. While attitudes toward the subject may have been formed many years previously, Ausubel (1968), Biehler (1971), and Blair, Jones, and Simpson (1967) agree that attitudes are not immutable and can be changed by skillful handling. Simonson (1976) reports a study in which attitudes were deliberately changed through use of the dissonance theory.

These studies indicate two things: that positive attitudes toward subject matter are desirable, and that attitudes can be changed in the desired direction by use of certain known techniques. If this is true, measurement of attitude, and especially measurement of attitude change during the term should be of value to the teacher. Johnson (1974) reminds us of the following:

> All learning has affective components, No matter what knowledge or skills a student masters, he will have feelings about the process and results of instruction. In mastering the skills of reading or in learning about history, a student develops feelings about reading and about history, as well as about learning and instruction, that will influence his behavior in the future. Because students' affective responses to school experiences influence future behavior, the development

> of positive affective reactions may be more important than the mastery of specific knowledge and skills. It does little good to teach a student to read if he ends up disliking reading and avoids it whenever possible. (Johnson in Walberg, 1974, p. 99)

Some authors (e.g., Ebel, 1972) would limit assessment to cognitive outcomes alone. Others (Johnson, 1974; Krathwohl, Bloom, & Masia, 1964) believe that affective components should be included in the evaluation model. Still other writers might wish to include assessment of some of the other outcomes enumerated above. It is apparent, however, that no practicable evaluation plan can include all of these components.

Model and Purpose

Model. The B-model is concerned with cognitive and affective outcomes. This is not intended to negate the importance of the other outcomes. It recognizes the fact that pupil gain scores in such nebulous areas as citizenship and moral character are extremely difficult to measure and are affected by many factors other than the instructional program. The model presented here is a systematic approach to monitoring class progress in the cognitive and affective areas periodically throughout the school term.

Other models for evaluating class progress measure accomplishment only at terminal points, e.g., the end of a chapter, the end of a unit, the end of a semester or school year. Because this model takes regular, frequent measurements of progress towards cognitive and affective outcomes, several desirable results may be achieved that are lacking with traditional models. First, the instructor will know at frequent, periodic intervals how his class is progressing toward accomplishment of his objectives for the term. Second, the instructor can accurately assess the amount of progress from one period to the next. Third, if the slope of the learning curve appears to flatten or to descend, the instructor can take remedial action at once. Fourth, the students themselves can observe the progress of their class as a whole. Fifth, if the model is used in the same kinds of classes with the same type of students, typical learning curves will become apparent. These can serve as standards of comparison for teachers and their supervisors and aid in the identification of effective teaching. The model will therefore serve three purposes:

1. to monitor class progress,
2. to motivate students,
3. to serve as a tool to assist in assessment of teaching.

Purpose. The present study provided for implementation of this model. The purpose of this study was to gain valuable information regarding the feasibility and practicability of the B-model for classroon use. This study was visualized as the first in a series of trials of the model in various settings to determine its potential utility for measuring gains in achievement and changes in attitudes of students over a period of time.

The B-Model

The B-model measures change in the level of pupil achievement by the use of multiple matrix sampling (for more information on this method see Appendix A, Multiple Matrix Sampling). The model extends beyond other designs for measuring class progress in two important respects. The B-model includes measurement of attitudinal as well as cognitive changes and includes more than just pre-test and post-test measures as recommended by Shoemaker (1977). It calls for testing at eight to ten periodic intervals throughout the school term, thus providing the teacher with valuable information to guide the instructional program.

The unique nature of the model with its multiple measurements taken during the term may be illustrated by contrasting it with a typical example of program evaluation made in the traditional way. A study reported by Leinhardt (1977) designed to evaluate a program of study included data from four different sources: standardized tests, questionnaires, videotapes, and student records. These measurements, however, were taken only in the fall and in the spring. This pattern of fall-spring measurement has been typical of previous attempts to monitor class performance or to evaluate program success.

The B-model to monitor class progress, with respect to student knowledge and attitude, would consist of observations made at equal intervals throughout the school term, using multiple matrix sampling techniques as described in Sirotnik (1974). Although there is no general agreement on what constitutes a learning curve for a given group (Hilgard & Bower, 1966), most educators would agree that given a set, or item domain, of test items designed to measure what a teacher is teaching, the percentage of items answered correctly by the teacher's students should increase over time. The amount of increase would be dependent on a number of factors, including intelligence, motivation, social conscience level, etc., of the students, and the effectiveness of the teacher.

The components of the model are the following:

1. A set of objectives for the particular subject matter area,
2. an item domain of test items which will measure these objectives,
3. a computer system which will randomly select items from the item domain to be used for the periodic tests,
4. a system that is both efficient and effective for producing, administering, and scoring the tests, and
5. the return of information on class progress to the instructor and to the students.

Shoemaker (1975) indicated the desirability for achievement tests derived from instructional programs by means of the item universe concept.

> An instructional program and its associated item universe are isomorphic. For every instructional program there exists one and only one item universe that is inseparable conceptually from it. The item universe is an operable definition of the instructional program. (pp. 128-129)

Shoemaker further asserts that the instructional program is the vehicle for providing the necessary skills to answer correctly all items in the item universe, not by teaching the correct responses to each item, but by teaching algorithms or concepts needed by students to respond appropriately. Most item universes, however, are so large as to be unmanageable and are therefore impossible to work with. A workable item universe can be found in the form of the "item domain". An item domain is a definable and enumerable subuniverse of items selected from the item universe in such a way that it includes every area in the item universe. Thus achievement, as measured by the item domain, will be equivalent to achievment measured by the item universe. An item domain for a given area might realistically include 500 to 2000 items.

Assessment of group progress toward accomplishment of the course objectives can be made by use of multiple matrix sampling in which the item domain is divided into small subtests and each subtest administered to a group of students sampled randomly from those participating in the instructional program. Because each student responds to only a small portion of the total number of items available, the testing program need not utilize a large amount of class time.

The B-model does not assume that classes should progress at equal rates. After each testing period, the means of cognitive items for each class are to be plotted as a curve of accumulated learning. From previous studies of such curves (Hilgard, 1956), it is assumed that the learning curves will show patterns similar to those in Figure 1.

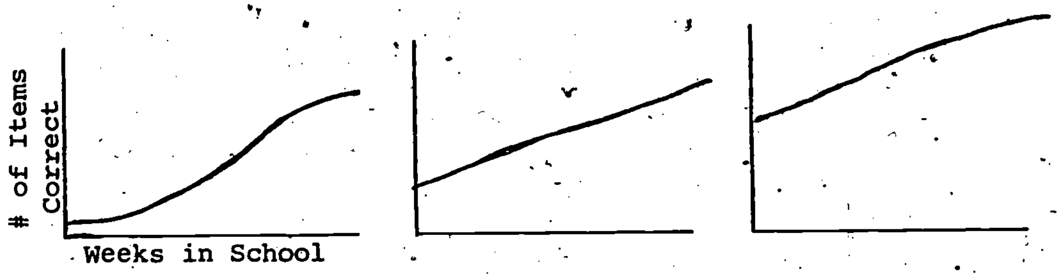

Figure 1

Examples of Expected Class

Learning Curves

It will be observed that not all classes begin or end at the same place on the scale, nor do all show the same rate of improvement.

This reflects the different abilities of the various groups and is to be expected. Each school system would have to develop its own set of learning curves to discover what kinds of patterns should be expected in various teaching positions.

The same considerations described above for cognitive items must be utilized in dealing with the affective domain. Affective objectives must be outlined clearly; items must be written to test students' attitudes; periodic multiple matrix sampling can be used to evaluate the students' position with regard to the objectives. In this way the B-model measures both cognitive and affective aspects of class performance. The goal in teaching is to increase the cognitive level of the students with regard to the subject matter and to improve, or at least to maintain, affective disposition toward the subject.

For an implementation of the B-model the logical starting point is the writing of a set of objectives that describe in detail all of the teacher's attitudinal and cognitive goals for the course. If several different teachers teach the same course, this should either be a cooperative project or the project should be assigned to one or a few teachers and carefully reviewed by all teachers who will be involved with the course. A process of revision of objectives should continue until all teachers can agree on the following:

1. The objectives listed are reasonable and desirable ones for the course in questions.
2. The objectives listed are (for the vast majority) the topics I intend to cover in the course. (The teacher may have some additional objectives not included on the list.)

A set of test items must be compiled that test each objective listed. These items are either composed after the objectives are written or collected from a pool of test items that may already be in existence for the course. There must be at least one test item for each objective; and no test item should refer to a topic not covered in the list of objectives. Ideally, several test items would be written for each objective.

Once the sizes of the item domains for both attitudinal and cognitive objectives have been determined, a computer program can be written (e.g. Barcikowski & Patterson, 1972) which is designed to select items at random from each domain and print tests. Each item, cognitive and attitudinal, is numbered and either typed on computer cards or stored on computer tape. The number of items to be chosen for each subtest from each domain will be determined by the size of the class, the sizes of the item domains, and the time available for testing. The number of subtests of each type to be printed will depend on the number of students placed in each subgroup.

The principal advantages in theory of the B-model for measuring progress of groups of students are the following:

1. Information that shows progress of students toward accomplishment of the objectives for the term is given to students and teachers on a regular, frequent basis.
2. Students and teachers can watch the curve of accumulated learning ascend as the students' knowledge increases. This should have a motivating effect on students and teachers alike.
3. Teachers can be alerted to attitude changes of the students.
4. Teachers can compare learning curves from one group to

another. They can then investigate the possible reasons for differences in learning curves from one term to the next.

5. Supervisors and administrators can use the learning curves to identify consistently outstanding teachers with the idea of attempting to determine possible causes for their consistent superiority.

The ultimate use of the B-model is to provide information available in no other convenient way that can be used for improvement of instruction. It is a systematic approach to measuring group progress.

It should be noted, however, that the B-model is not considered suitable for all types of classes. While all classes have cognitive and affective objectives, not all of these lend themselves to measurement by short-answer objective tests. Some types of objectives require too much time for measurement in a multiple matrix sampling plan. Subjects such as history, mathematics, science, and certain courses in English are the kinds considered suitable for measurement with the B-model.

Problem and Methodology

The problem to be answered in the study is the following:

Is the B-model practical and feasible, for measuring the achievement and attitudes of students over time?

Classes in which the model was tested were five sections of EDRE 501, Introduction to Research Methods, offered at the Ohio University during the fall quarter, 1978, taught by three different instructors. Two of the classes were offered on the main campus, one in the evening, one in the morning, the other three classes were offered on three branch campuses in the evening. All classes were offered once a week and met for three hours.

The study was originally designed to have five different instructors, however, two instructors left the University for reasons unrelated to the study. The remaining three instructors were assigned so that one instructor taught three classes (one branch and the two classes on campus) and the other two instructors each taught one class off campus. One instructor was a male full professor who had taught this course, or a similar course, at least once each year for the past twenty years. Another instructor was a male assistant professor who had taught this course several times over the past three years. The third instructor was a female who had had ten years experience teaching at the high school level, and who had finished all of her course work towards her doctorate in Educational Administration. The latter instructor had taken this course as a student, but had never taught it. All of the instructors knew that they would not be identified in the report of the study.

The following plan of procedures was followed:

1. A list of objectives for the course, in both the cognitive and affective domains, was prepared.
2. An item domain that was congruent to the list of objectives was assembled.
3. An instrument to measure attitudinal objectives was compiled and piloted.
4. The number of items from each item domain that needed to be sampled for each subtest were determined using procedures

described in Sirotnik (1974).

5. A computer program was written that selected and printed the requisite number of copies of each subtest.

6. A signed consent to participate in this study was obtained from each student in all five classes.

7. Demographic data from students in each EDRE 501 class was collected using the form shown in Appendix B.

8. Tests were administered at weekly intervals to the students in the EDRE 501 classes.

9. These tests were scored and class means were determined each week.

10. Close account was kept of all time spent writing objectives, writing test items, administering and scoring tests.

11. The mean for each class was plotted on a separate graph and copies were distributed to the instructors. Each subsequent mean was plotted on the same graph to indicate class progress.

Criteria for Success

The factors that distinguish the B-model from the other models are the simultaneous use of pupil-gain measures and attitude change measures, economy of teacher and student time, and provision of helpful information throughout the school term. It was decided beforehand that the model would be judged successful if the following criteria were met:

1. The multiple matrix sampling technique must be able to measure changes in student achievement and attitude. This would be shown by the curves on the graphs of achievement

and attitude for each class. The differences in means between times would be tested for significance at the .05 level using a multivariate repeated measures design.

2. The expenditure of classroom time for testing must not be judged too high by the instructors. The exact amount of time involved for giving instructions and for administering test items would be recorded. The determination of whether or not this time is excessive will be a subjective judgment based on the instructors' opinions.

3. The instructors must find that the information on achievement and attitude contributed to their understanding of the progress of their classes. The determination of the worth of this information would be a subjective judgment by the instructors. Each instructor would be asked to respond to questions about this using a structured interview (Appendix C).

Results

Construction of the Cognitive and and Attitude Domains

A list of 165 cognitive and 14 attitudinal objectives were compiled and agreed upon by the instructors. An item domain was then established which was congruent with the objectives, and which consisted of 238 items measuring student achievement and 58 items measuring student attitude. Of the 238 achievement items, 124 were judged by the instructors to be knowledge items, 75 were judged to be understanding items, and 38 were judged to be application items. All of these achievement items were taken from past tests for this course. The Kuder-Richardson 20 reliabilities of these past tests ranged from .63 to .88, on tests composed of from 25 to 50 items.

The attitude items were constructed based on information gathered from three EDRE 501 classes given during the 1978 summer sessions at Ohio University. Initially twenty-one students in one EDRE 501 class were asked to respond to eight open-ended questions concerning their likes and dislikes towards educational research. From the responses to these open-ended questions a list of 100 attitude items was compiled. These attitude items were then tried on a group of 19 students in a second EDRE 501 class, and based on their responses items were modified or deleted. A revised list of 70 items were then given to twenty-two students in a third EDRE 501 class. The 58 items for the final attitude instrument were selected because they yielded mean differences between groups who scored high and low (total attitude score plus or minus .5 standard deviation above or below the mean) of at least .2 of a standard deviation and had a correlation with the total test score of at least .25. The final attitude instrument had 35 positively worded items and 23 negatively worded items.

Classes

A brief description of the students who enrolled in the five classes in this study is shown in Table 1. The information in Table 1 was arrived at from the background information sheet in Appendix B. In Table 1 it can be seen that the classes differed considerably on whether the students were full time (registered for fifteen or more hours) or part-time (registered for less than fifteen hours). Classes 1, 2, and 5 were composed of primarily part-time students;

Table 1

Frequency Counts and Percentages[a] of Students in Various Categories Across the Five Classes

Category	Class					
	1[b]	2	3[b,c]	4[b,c]	5	Overall
Type of Enrollment						
Full time	4(27)	1(9)	12(60)	11(100)	0(0)	28(43)
Part time	11(73)	10(91)	8(40)	0(0)	8(100)	37(57)
Undergraduate Degree in Education						
Yes	10(67)	10(91)	14(70)	8(73)	5(63)	47(72)
No	5(33)	1(9)	6(30)	3(27)	3(37)	18(28)
Age						
18-25	3(2)	5(45)	10(50)	3(27)	3(38)	24(37)
26-35	9(60)	5(45)	8(40)	7(64)	4(50)	33(51)
36-45	2(13)	1(10)	2(10)	1(9)	1(12)	7(11)
46-55	1(7)	0(0)	0(0)	0(0)	0(0)	1(1)
Sex						
Male	5(33)	1(9)	6(30)	6(55)	1(12)	19(29)
Female	10(67)	10(91)	14(70)	5(45)	7(88)	46(71)
Class Size						
Class Size	15	11	20	11	8	65

[a] Percentages are in parentheses.

[b] Taught by one instructor.

[c] Taught on campus

class 3 had eleven (60%) full-time and eight (40%) part-time students; and class 4 had all full-time students. All of the classes were composed primarily of students who had received their undergraduate degrees in Education. Across all of the classes the students were primarily (88%) in the 18-35 age range, however, classes 1, 4 and 5 had slightly older (26-55) students, and classes 2 and 3 had about half of their students in the younger (18-25) age range. All of the classes were primarily made-up of females, except for class 4 which was approximately evenly split between males (55%) and females (45%).

Classes 1, 2 and 5 were offered in the branch campuses, and classes 3 (evening) and 4 (morning) were offered on the main campus. From Table 1 the main distinction between the branch and main campus students was that most (68%) of the main campus students were enrolled full-time, while most (85%) of the off-campus students were enrolled part time. One instructor taught classes 1, 3 and 4.

Test Size and Time

As can be seen in Table 1 the classes consisted of different numbers of students. Following the multiple matrix sampling procedure (See Appendix A) this meant that the size of the test taken in each class was dependent on the number of students in the class. In all cases the total item domain both cognitive and attitude were used. This meant that in Class 1 with 15 students each week, each student took a 15 or 16 item achievement test, and a 3 or 4 item attitude test. The approximate size of each test taken each week from each item domain is shown in Table 2 along with the average time spent taking each test. As can be seen in Table 2 the average testing times ranged from 13.6 minutes to 17.7 minutes. Therefore,

Table 2

Approximate Number of Items Taken Each Week in the Five Classes and the Average Time Spent in Testing

	Class				
	1	2	3	4	5
	Approximate Number of Items (Average Time in Minutes)				
Achievement	16	22	12	22	30
Attitude	4	5	3	5	7
Total	20(14.3)	27(13.6)	15(13.6)	27(14.7)	37(17.7)

for future testings with item domains similar in size to the ones used here, and with classes larger than 10 students, one might plan on a test period of approximately 15 minutes.

In order to allow for absentees it was necessary to construct a series of 36 test groupings. This series allowed the total item domains to be tested each week and also controlled so that no student in Classes 1 through 4 ever took the same item twice. In Class 5 there were only eight students and 10 testings, therefore, each person took items they had seen before (but at different times) during the last two testings. Tests were distributed at random during the first session, but then students and tests were kept track of to ensure that no student took the same test.

Approximately 40 hours were spent collecting, classifying, and collating the items for the cognitive domain. Another eight hours were spent revising and correcting these items. Approximately six hours were spent writing the initial 100 attitudinal items, with an additional seven hours spent revising the attitudinal items and objectives, and testing these items. Therefore, approximately 61 hours were spent preparing the item domains. The computer time required to prepare and score the tests each week was 25.4 seconds; the human time needed each week to pull the tests apart, distribute them, and have them scored was one hour and 38 minutes.

Analysis of Class Means

The estimates of the class means for the cognitive items are plotted in Figure 2 and the estimates of the class means for the attitude items are plotted in Figure 3. The overall multivatriate repeated measures analysis of these class means is shown in Table 3

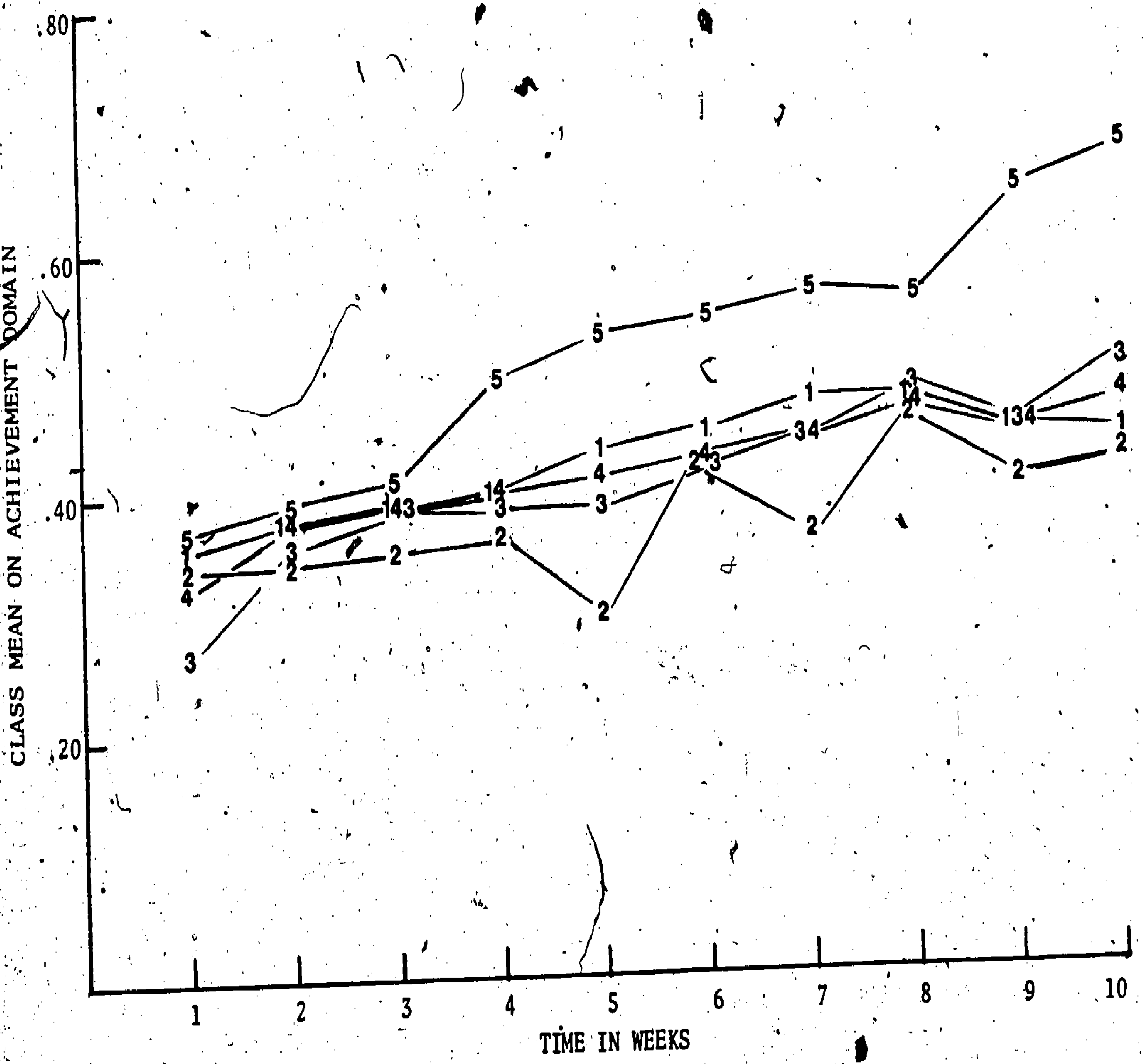

Figure 2

The Changes in Class Mean on the Achievement Domain for the Five Classes in This Study

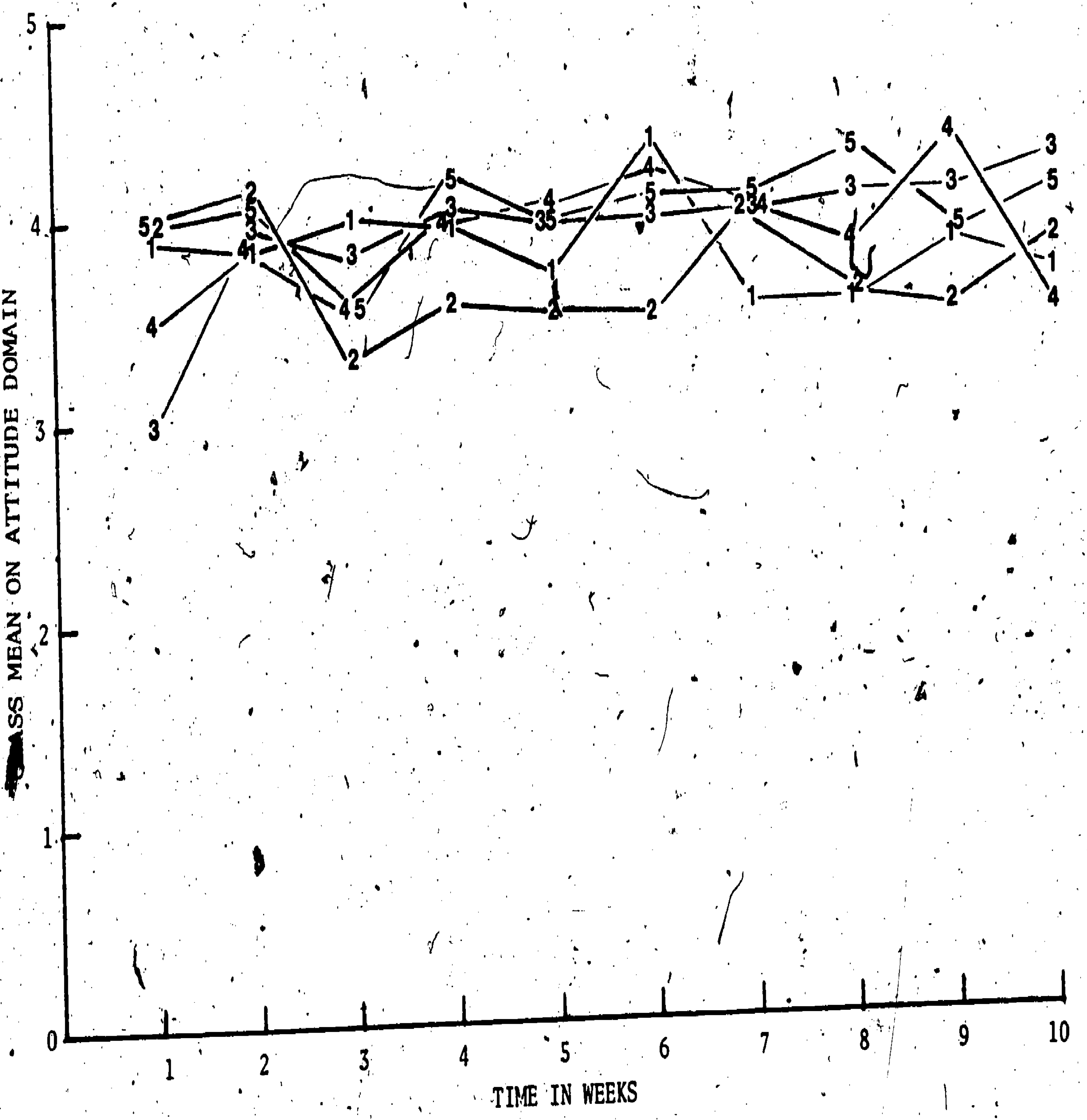

Figure 3

The Changes in Class Mean on the Attitude Domain
for the Five Classes in This Study

Table 3

Overall Multivariate Repeated Measures Analysis
with all Five Classes

Source of Variation	Sums of Squares and Products Achievement	Sums of Squares and Products Attitude	Multivariate Df	Multivariate Wilks' Lambda	Multivariate F(Probability)	Univariate Df	Univariate F(Probability) Achievement	Univariate F(Probability) Attitude
Time	.1366	.1503	18	.23	4.13*(.0001)	9	10.56*(.0001)	.91(.5298)
	.1503	.6002						
Classes	.1311	.2522	8	.26	8.41*(.0001)	4	22.81*(.0001)	2.33(.0742)
	.2522	.6865						
Error	.0517	.0550	70			36		
	.0550	2.65						

*Significant at $p < .05$

Table 4

Multivariate Repeated Measures Trend Analysis Over Time with all Five Classes

Source of Variation	Sums of Squares and Products Achievement Attitude	Multivariate Df	 Wilks' Lambda	 F(Probability)	Univariate Df	F(Probability) Achievement	 Attitude
Linear Trend	$\begin{bmatrix} .1299 & .1203 \\ .1203 & .1115 \end{bmatrix}$	2,35	.28	44.02*(.0001)	1	90.52*(.0001)	1.52(.2262)
Other	$\begin{bmatrix} .0067 & .0300 \\ .0300 & .4887 \end{bmatrix}$	16,70	.76	.64 (.8401)	1	.58.(.7859)	.83(.5815)
Error	$\begin{bmatrix} .0517 & .0550 \\ .0550 & 2.65 \end{bmatrix}$				36		

*Significant at ρ < .05

with the trend analysis over time shown in Table 4. The overall multivariate analysis of the means for the three instructors, using the unweighted average of the means of the three classes taught by one instructor, is shown in Table 4 with the trend analysis over time shown in Table 5.

In Figure 2 the learning curves for the classes are linear with a positive slope, and there appears to be class differences with respect to achievement. In Figure 3 the attitude means show no trend over time and no differences between the classes. These observations are supported by the results shown in Tables 3 and 4. In Table 3 a multivariate significant difference is found between the class means over time ($F = 4.13$, $\rho < .0001$) and most of this difference is due to achievement ($F = 10.56$, $\rho < .0001$) and not attitude ($F = .91$, $\rho < .5298$). In Table 4 a multivariate linear trend over time is indicated ($F = 44.02$, $\rho < .0001$), and the linear trend is found over the achievement means ($F = 90.52$, $\rho < .0001$) but not over the attitude means ($F = 1.52$, $\rho < .2262$). The results in Table 3 also indicate a multivariate significant difference among classes ($F = 8.41$, $\rho < .0001$) with this difference primarily due to achievement ($F = 22.81$, $\rho < .0001$) and not attitude ($F = 2.33$, $\rho < .0742$).

The statistical analyses in Tables 5 and 6 are the same as those shown in Tables 3 and 4 except that the comparisons were made with respect to instructors and not classes. In these tables the class means from the classes taught by one instructor were averaged to represent him. The results are the same as those reported in Tables 3 and 4 with the exception that the overall multivariate

Table 5

Overall Multivariate Repeated Measures Analysis
With Three Classes[a]

	Sums of Squares and Products		Multivariate			Univariate		
Source of Variation	Achievement	Attitude	Df	Wilks' Lambda	F(Probability)	Df	F(Probability) Achievement	Attitude
Time	.0967	.0441	18	.22	2.16*(.0257)	9	4.71*(.0026)	.7812(.6363)
	.0441	.0364						
Instructors	.1252	.2648	4	.21	10.04*(.0001)	2	27.43*(.0001)	5.83* (.0112)
	.2648	.6034						
Error	.0411	.0059	34			18		
	.0059	.9313						

[a]the scores for the three classes taught by the same instructor were averaged to form the third class.

*Significant at $p < .05$

Table 6

Multivariate Repeated Measures Trend Analysis Over Time with Three Classes

	Sums of Squares and Products		Multivariate			Univariate		
							F(Probability)	
Source of Variation	Achievement	Attitude	Df	Wilks' Lambda	F(Probability)	Df	Achievement	Attitude
Linear Trend	.0923	.0396	2,17	.31	19.16*(.0001)	1	40.43*(.0923)	.33(.5735)
	.0396	.0170						
Other	.0045	.0046	16,34	.66	.50 (.9324)	8	.24 (.9760)	.84(.5821)
	.0046	.3468						
Error	.0411	.0059				18		
	.0059	.9313						

*Significant at $p < .05$

significant difference between the instructors ($F = 10.04$, $\rho < .0001$) appears to have been due to both achievement ($F = 27.43$, $\rho < .0001$) and attitude ($F = 5.83$, $\rho < .0112$). This result led to the plotting of the three instructor's classes attitude means in Figure 4. The plot of the attitude means in Figure 4 indicates that there are interactions in the data making the overall test difficult to interpret. However,Class 2 does have the lowest pattern of attitude and did finish the class with a lower attitude mean than it started with. Classes 5 and C had a higher pattern of mean attitudes and both finished higher than they started.

Instructor Opinions

The information discussed in this section is based on the structured interview questions found in Appendix C. In response to questions #1, #2 and #3 concerning worthwhile achievement and attitude information, the instructors indicated that they found the information interesting but that they made no use of it. In response to questions #4 and #5 concerning information to students, the instructors indicated that while the students showed some interest in the class's progress, they were primarily interested in their own individual progress. In response to question #6 two instructors thought that the testing periods did not require an excessive amount of time, and one instructor indicated that the testing time added up to one class period, and that was substantial for a class that meets only once a week. The instructors responses to questions #7 and #8 concerning information gained from this study, indicated that by itself the information was not of value to them in their teaching, but that if the class information could be considered with respect to the other classes, or

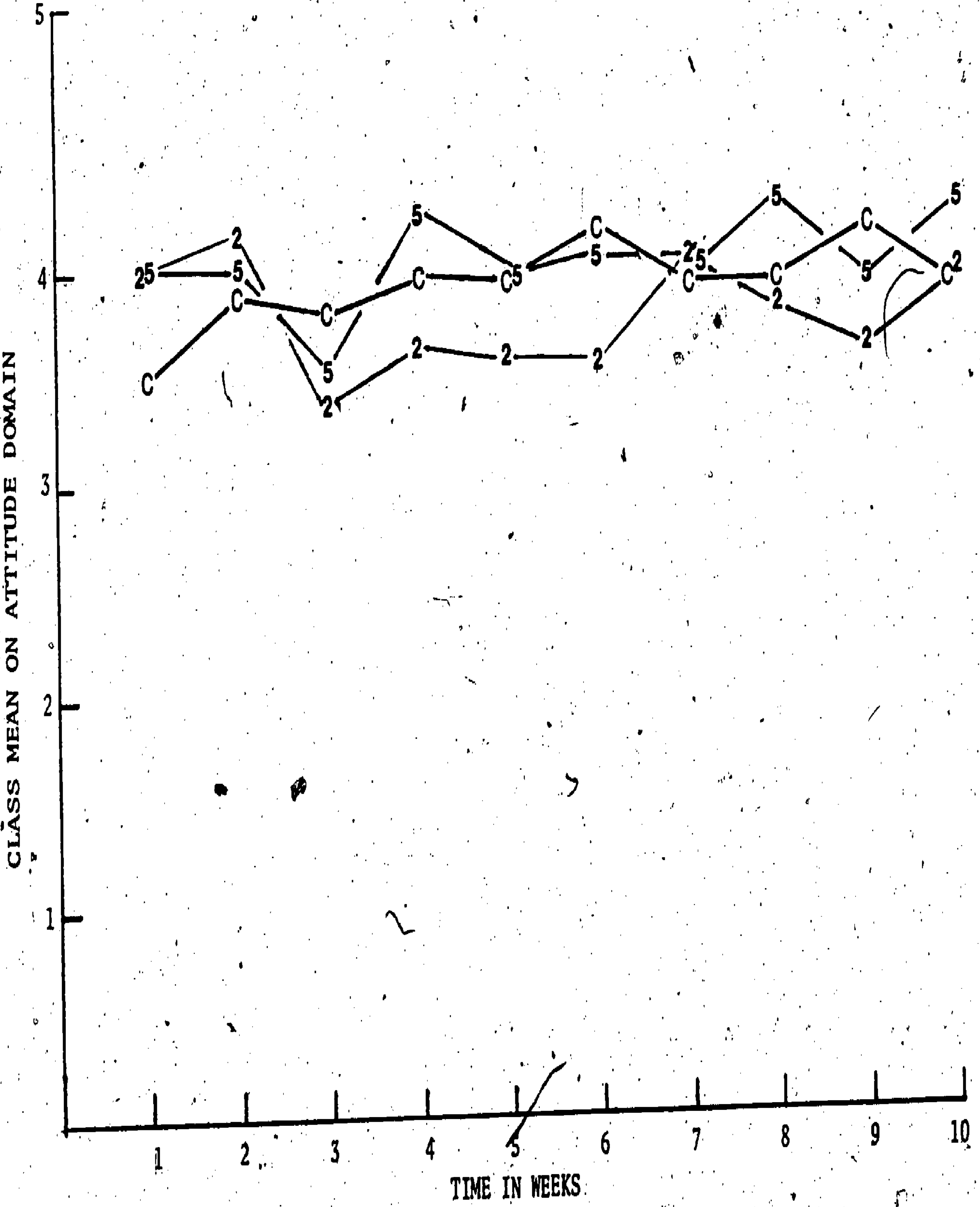

Figure 4

The Changes in Class Mean on the Attitude Domain for the Three Instructor's: Classes-C is the Combined mean (Average) for One Instructor with respect to normative information, it might serve to motivate them to improve their teaching.

Discussion

Although a good deal of care was put into implementing this study of the practicality and feasibility of the B-model, the results can only be considered as exploratory in considering the model's full potential. This was the first implementation of this model and it was done on a small scale with only five classes and three teachers. However, some of the results, particularly those in the measurement of the cognitive domain may be considered as particularly encouraging. If one reconsiders the distributions of cognitive class mean estimates over time, shown in Figure 2, it is interesting to note that although the classes start in different positions on the first testing, by the second testing the classes establish a pattern that seems to be reflecting teacher differences. Here the teacher with Classes 1, 3, and 4 has established the median pattern of achievement; the teacher in Class 5 the highest pattern of achievement; and the teacher in Class 2 to lowest pattern of achievement. What is of interest is that the classes taught by one instructor had mean achievement that was so homogeneous and yet distinguishable from the achievement of the other two classes. This pattern ~~is~~ most interesting when one considers the variety of background information exhibited by Classes 1, 3 and 4 in Table 1. Wouldn't this study be fascinating as an experiment where subjects were randomly assigned to classrooms? Is it the teacher in Class 5, the students, or something that this teacher is doing that is causing the high achievement (or is it simply the fact that these students took more items)? What is happening in

Classes 1, 2, 3, and 4? Although this data may simply be an artifact of its small scale, the achievement results certainly encourage its implementation in a larger scale.

The results of the instructor opinions were not judged to be encouraging for use of the B-model. But this seemed to be caused by the way the results were presented to each instructor, with no perspective. That is, the instructors did not know if they were doing well or not -- they had no criteria on which to make a judgment as to their classes progress. This is an indication that in future use of the model more effort should be put into providing the instructors with comparative information - although this may require a study with the same instructors over several years. (What might happen to class achievement if an experimental study were conducted where instructors were shown learning curves with their classes above or below the norm?)

The measurement of student attitude towards educational research failed to indicate any reliable teacher difference or differences over time. The results in Figures 3 and 4 indicate that, at least on the instrument constructed for this study, the teachers had high positive attitudes towards educational research, and that these attitudes did not appear to be strongly affected by their teachers.

Conclusion

In their recent book on school effectiveness Madaus, Aivasian, and Kellaghan (1980) indicate that in the United States educators have traditionally used inappropriate measures (i.e., standardized tests) to study the effects of schooling. They indicate that these inappropriate measures are at least partially responsible for the results shown in many studies (e.g., Circirelli, et al., 1969; Coleman, et al., 1966; Glass, et al., 1970; Jensen, 1969) that schools (teachers) have no strong effects on student achievement beyond that accounted for by measures of social class and home background. The results presented here indicate that the B-model may provide researchers and school administrators with a sensitive measurement approach, which is economical in terms of teacher and student time, with which to measure pupil progress, and perhaps teacher effectiveness, throughout the school year.

References

Ausubel, D. B. Educational psychology: A cognitive view. New York: Holt, Rinehart, and Winston, Inc., 1968.

Barcikowski, R. S., & Patterson, J. L. A computer program for randomly selecting test items from an item population. Educational and Psychological Measurement, 1972, 32, 795-798.

Biehler, R. F. Psychology applied to teaching. Boston: Houghton-Mifflin Co., 1971.

Blair, G., Jones, R., & Simpson, R. Educational psychology. New York: MacMillan Co., 1967.

Cicirelli, V. G., et al. The impact of Head Start. An evaluation of the effects of Head Start on children's cognitive and affective development. Study by Westinghouse Learning Corporation and Ohio University. Washington, D.C.: Office of Economic Opportunities, 1969.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. Equality of educational opportunity. Washington, D.C.: Office of Education, U. S. Department of Health, Education and Welfare.

Cronbach, L. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1972.

Ebel, R. Declining scores: A conservative explanation. Phi Delta Kappan, 1976, 58, 306-310.

French, J. W. Aptitude and interest score patterns related to satisfaction with college major field. Educational and Psychological Measurement, 1961, 21, 2, 287-294.

Gallup, G. Ninth annual Gallup poll of public attitudes toward education. Phi Delta Kappan, 1977, 59, 33-48.

Glass, G. V., et al. Data analysis of the 1968-69 survey of compensatory education, Title I, Final Report No. OEG8-8-961860 4003-(058), Washington, D. C.: U. S. Office of Education, 1970.

Hilgard, E. R. Theories of learning, 2nd ed. New York: Appleton-Century-Crofts, Inc., 1956.

Hilgard, E. R., & Bower, G. H. Theories of learning, 3rd ed. New York: Appleton-Century-Crofts, Inc., 1966.

Johnson, D. W. Affective outcomes. In H. S. Walberg (Ed.) Evaluating educational performance. Berkeley, California: McCutchan Press, 1974, 99-112.

Krathwohl, D. R., Bloom, B. U., and Masia, B. B. Taxonomy of educational objectives, Handbook II: Affective domain. New York: David McKay Co., Inc., 1964.

Leinhardt, G. Program evaluation: An empirical study of individualized instruction. American Educational Research Journal, 1977, 14, 277-293.

Madaus, G. F., Aivasian, P. W., & Kellaghan, T. School effectiveness. New York: McGraw-Hill, 1980.

Porter, J. The virtues of a state assessment program. Phi Delta Kappan, 1976, 57, 667-668.

Shoemaker, D. M. Principles and procedures of multiple matrix sampling. Cambridge, Massachusetts: Ballinger, 1973.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-147.

Shoemaker, D. M. The contribution of multiple matrix sampling to evaluating teacher effectiveness. In Borich, G. D. (ed.) The appraisal of teaching: Concepts and process. Reading, Massachusetts: Addison-Wesley Publishing Co., 1977, 292-300.

Simonson, M. R. Attitude change and achievement: Dissonance theory in education. Journal of Educational Research, 1976, 21, 163-169.

Sirotnik, K. Introduction to matrix sampling for the practitioner. In Popham, W. J. (ed.) Evaluation in education. Berkeley, Calif.: McCutcheon, 1974. (Also available as a separate paperback, same publisher, same year.)

Worthern, B., & Sanders, J. Educational evaluation: Theory and practice. Worthington, Ohio: Charles A. Jones, 1973.

Reference notes

Barcikowski, R., & Upp, C. *A model system for the evaluation of teacher effectiveness.* Unpublished manuscript, 1978. (Available from Robert Barcikowski, Ohio University, Athens, Ohio, 45701.)

APPENDIX A

MULTIPLE MATRIX SAMPLING

Multiple matrix sampling is a method of collecting group data with the expenditure of a very small amount of time and money as compared to the traditional census method of collecting data. In the census method of testing, all items are administered to all students. For example, 25 arithmetic items might be administered to a class of 30 pupils, with each pupil being tested on all 25 items. Individual achievement data may be collected in this manner and group statistics may be derived from the individual data. Note that 750 items (25 items x 30 pupils) would need to be scored for the example given. If individual data are not needed multiple matrix sampling can greatly reduce the number of items to be scored, thus providing economy of pupil and teacher time.

To apply a procedure of multiple matrix sampling the group of test items is divided into subtests, and the group of examinees is divided into subgroups of examinees. This is done by a procedure of randomization in both cases. For a small number of items and examinees as given in the example above, this can be done by using a table of random digits. If the decision has been made to reduce testing time to one-fifth, items and examinees are randomly dividied into five groups. Students are assigned sequential numbers beginning with 1. Test items are assigned sequential numbers beginning with 1. The table of random digits is then used to select the items for each subgroup. For the example above, the random arrangement would give five subtests, which might contain the following items:

Subtest 1 - items 2, 8, 16, 5, 17

Subtest 2 - items 4, 14, 7, 20, 15

Subtest 3 - items 10, 19, 25, 6, 18

Subtest 4 - items 12, 22, 1, 13, 3

Subtest 5 - items 9, 11, 21, 23, 24

A random arrangement of 30 students into five groups might produce the following arrangement:

Subgroup 1 - students 7, 18, 28, 4, 5, 20

Subgroup 2 - students 3, 22, 27, 2, 17, 21

Subgroup 3 - students 29, 13, 11, 14, 23, 26

Subgroup 4 - students 6, 19, 15, 30, 24, 9

Subgroup 5 - students 1, 8, 10, 12, 16, 25

Subgroup 1 would then be given subtest 1, subgroup 2 would be given subtest 2, etc. Each subgroup of students is given a different group of items, a fraction of the size of the original test. A mean score for each subgroup is computed (the number of items answered correctly by students in each subgroup divided by the total number of possible responses for that subgroup). From these means the mean of the entire group is computed. This mean of the subgroup means is an unbiased estimate of the true mean of the group and will correspond very closely with the mean that would be determined by administering every test item to every student. Note that in the example given only one-fifth as much time for test administration was required and only one-fifth as many items (30 students x five items each) need to be scored. Sirotnik (1974, p. 461) and Shoemaker (1973, p. 5) both indicate the accuracy of multiple matrix sampling as an estimator of group means and both give excellent examples.

When the aim of testing is to measure the degree of accomplishment of the objectives for a complete course of study, the test item pool might easily consist of several hundred or even two or three thousand test questions. This would, of course, necessitate the assignment of more than five or six items to each subgroup of pupils. The exact number of items to be assigned is determined by the size of the item pool, the number of students to be tested and the degree of accuracy desired. If the item pool contains fewer than 500 test items, every question should be answered by at least one subgroup of pupils (Sirotnik, 1974, p. 467). If the item population is larger than 500, it can be considered as of infinite size and sampled randomly to obtain subtests of appropriate size. A more complete discussion of "appropriate size" may be found in Sirotnik (1974) or Shoemaker (1973).

APPENDIX B

DEMOGRAPHIC DATA

STRUCTURED INTERVIEW

1. Did you receive any worthwhile information on the achievement of your class during the progress of the study?

2. Did you receive any helpful information on the attitudes of your class during the course of the study?

3. Would you have received this information without the study? If so, how? In what form?

4. Do you feel that your students were interested in the shape of the learning curve as it developed?

5. Do you think that knowledge of the class progress was beneficial or harmful to the class?

6. Did the testing periods require an excessive amount of class time?

7. Did the information you received assist you in understanding the progress of your class?

8. Would you like to continue to receive this kind of information about your classes?

APPENDIX C

STRUCTURED INTERVIEW FORM

BACKGROUND INFORMATION

To help in the analysis of data for the study you are participating in, some background information is required. Please answer the questions below. You need not sign your name.

1. Full or part-time student? ____________________

2. If employed, where? ____________________

3. What position do you hold? ____________________

4. What position do you hope to hold after you complete your studies?

____

5. If you are a teacher, how many years have you taught? __________

6. Major field? ____________________

7. Age:
    0-17 __________
    18-25 __________
    26-35 __________
    36-45 __________
    46-55 __________
    56-? __________

8. Sex:
    Male __________
    Female __________