DOCUMENT RESUME

ED 083 301 TM 003 266

AUTHOR Semb, George
TITLE Research Strategies in Higher Education.
INSTITUTION Kansas Univ., Lawrence. Dept. of Human Development.
PUB DATE Aug 73
NOTE 20p.; Paper presented at annual convention of American Psychological Association (Quebec, Canada, August 1973)
AVAILABLE FROM George Semb, Department of Human Development, University of Kansas, Lawrence, Kansas 66044

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Achievement Tests; College Teachers; *Effective Teaching; *Evaluation Criteria; Higher Education; *Research Methodology; Speeches; *Teacher Evaluation; Testing

ABSTRACT

The present paper outlines two alternative strategies for evaluating teaching effectiveness. These are: (1) within-subject reversal designs, and (2) multiple baseline testing procedures. Each design is discussed in terms of its application to research problems in higher education. In reversal designs, the student is exposed to different teaching procedures during successive phases of a course. Changes in performance between treatments are analyzed, either on the basis of group averages or in terms of individual performances. The reversal design becomes even more powerful if a second group of students goes through the treatments in opposite order, thus counterbalancing the two groups for possible changes in difficulty across conditions. In the multiple baseline testing procedure, students are given a comprehensive examination before instruction begins, and at the end of each successive phase of the course. This allows the instructor to demonstrate that changes in performance are functionally related to specific teaching procedures introduced during each phase. Furthermore, it provides a continuous baseline measure over material that has been trained. Percentage gains over baseline levels can be used to measure differential effects of different teaching procedures. Similar to the reversal design, the multiple baseline design allows the researcher to make statements about the effects of each procedure on individual students. (Author/DB)

RESEARCH STRATEGIES IN HIGHER EDUCATION[1]

George Semb

Department of Human Development

University of Kansas

Paper presented at American Psychological Association Convention, Montreal, Quebec, Canada, August, 1973

RESEARCH STRATEGIES IN HIGHER EDUCATION

George Semb

Department of Human Development

University of Kansas

ABSTRACT

Researchers in higher education have traditionally used control-group/experimental-group techniques to compare different teaching procedures. Such experimental designs, however, frequently fail to account for individual differences in student performance and student preference. The present paper outlines two alternative strategies for evaluating teaching effectiveness. These include: a) within-subject reversal designs; and b) multiple baseline testing procedures. Each design will be discussed in terms of its application to research problems in higher education.

In reversal designs, the student is exposed to different teaching procedures during successive phases of a course. Changes in performance between treatments are analyzed either on the basis of group averages, or in terms of individual performances. The reversal design becomes even more powerful if a second group of students goes through the treatments in opposite order, thus counterbalancing the two groups for possible changes in difficulty across conditions.

In the multiple baseline testing procedure, students are given a comprehensive examination before instruction begins, and at the end of each successive phase of the course. This allows the instructor to

demonstrate that changes in performance are functionally related to specific teaching procedures introduced during each phase. Furthermore, it provides a continuous baseline measure over material which has not been trained, and a retention measure over material that has been trained. Percentage gains over baseline levels can be used to measure differential effects of different teaching procedures. Similar to the reversal design, the multiple baseline design allows the researcher to make statements about the effects of each procedure on individual students.

In 1968, Dubin and Taveggia (1968) published a comprehensive, comparative analysis of several different teaching methods - lecture, lecture-discussion, tutorial and independent self-study. Rather than accept authors' conclusions, they reanalyzed the data from some 91 studies conducted between 1925 and 1965. The results? There were no measurable differences among truly distinctive methods of instruction when evaluated by student performance on midterm and final examinations. Surprising? Discouraging? Expected? It all depends.

Perhaps midterm and final examinations do not, as some have argued, measure what was <u>really</u> taught, which places their existence in question. Perhaps midterms and finals are characterized by too much variability to discriminate even large effects. Perhaps, as Hilgard (in Dubin and Taveggia, 1968) has noted, the way in which course materials are programmed, usually through a textbook from which test items are derived, are so powerful that they override differences in teaching. Perhaps, too, the problem is the way and rigor with which teaching methods have been analyzed and evaluated. That is, the inability to find differences in the effectiveness of different teaching technologies may lie in the experimental strategies used to assess those techniques. Most likely, there is no single cause, but rather a multiple-choice combination of all of the above. The purpose of this presentation is to outline some alternative strategies for research in college teaching which place greater emphasis on individual student performance and preference, and which have shown that some variables do make a difference.

Static Group Designs

Most college teaching research has used static group experimental designs to compare different teaching methods. Typically, one group of students is exposed to teaching method A and a second group to teaching method B. The two groups (or more) are compared on the basis of performance on midterm exams, a final and/or student course ratings. Differences between groups are typically assessed using any of a variety of statistical tests to determine if they are significant at some predetermined probability level. Significant is an unfortunate word because it is sometimes used to imply important. For example, the only statistically significant result reported by Dubin and Taveggia (1968) was that students who studied course material performed better than a group of control students who did not take the course. Although this is an appropriate control procedure, it is hardly an important source of information for an analysis of effective college teaching procedures.

This is not meant to deny the usefulness of statistics as a tool, but rather to suggest that researchers need not be slaves to them. Group statistics are most useful when we ask actuarial questions of the form - what percentage of students prefer X, Y or Z, or what percentage of students withdraw from a given course. They are also valuable when we attempt to standardize test materials by determining what proportion of students answer an item or items correctly following a standard method of instruction. However, when we use static group designs to compare student performances or preferences, we may indeed find that differences among students far outweigh the effects generated by different teaching procedures.

Single Subject Analysis

In the present paper, we will describe and illustrate the use of two major within-subject or within-group experimental designs, ones which I think will enable us to achieve a more precise analysis of college teaching techniques. Single subject analysis typically involves monitoring a student's behavior frequently throughout the term, say ten to thirty times in comparison with traditional designs which assess student performance and/or preference infrequently. We have Keller's (1968) system of personalized instruction to thank for small units of material which enable researchers to measure student behavior at frequent intervals.

In recent years, personalized systems of instruction have received a great deal of popular and empirical support. Contrary to the history of college teaching research portrayed by Dubin and Taveggia (1968), several investigators (Born, Gledhill and Davis, 1972; McMichael and Corey, 1969; Sheppard and MacDermot, 1970) have demonstrated that personalized instruction produces significantly better examination performance and higher student ratings than more traditional, lecture-discussion methods. It is not my purpose here to advocate or criticize personalized instruction, but rather to borrow from it some features which may help us better evaluate our teaching methods.

Small Units of Material

One salient feature of personalized and programmed instruction is the division of course materials into several small parts or units. In our introductory child development course at the University of Kansas (which we teach by personalized instruction), we have sixteen unit

quizzes and a final examination which is analyzed in 16 corresponding parts. In addition to enabling a more precise analysis of student behavior, the use of short units also improves test performance (Semb, 1973). When students were required to study material for four units combined and then tested on an achievement test over the material, they performed about 20% worse than when they studied and were tested over four units separately. Thus, there is a data base as well as an experimental rationale for using relatively short units of material.

Within-Subject Designs - Rationale

The use of within-subject and within-group experimental designs in college teaching research is relatively new. Campbell and Stanley (1963) refer to these designs as quasi-experimental time series analyses. They have been used widely in operant (behavior analysis) research with both animals and children. The two major designs we will consider in the present paper are the ABA reversal design and the multiple baseline design. In both, the same subject or group of subjects are exposed to two or more treatment conditions in successive order. Thus, one major source of variability characteristic of static group designs, between subject variability, is eliminated. Another obvious advantage of single subject analysis is the emphasis placed on the individual student.

Reversal Designs

In the typical ABA reversal design, the behavior of interest (e.g., student performance or preference) is measured frequently over time to establish a baseline as a basis for forecasting what the level of the behavior will be in the future (Risley and Wolf, 1972). Thus, as Fig. 1 shows, one can analyze the trend of the data, a feature

Insert Fig. 1 about here

which is just as important as the absolute magnitude of the effect. For example, Fig. 2 shows the same magnitude of effect as Fig. 1,

Insert Fig. 2 about here

but the trend is obviously different. Given the increasing level of the behavior in Condition A (Baseline), it would be difficult to argue that the increase observed during B would not have occurred had the experimental condition not been introduced.

To continue our discussion of the reversal design, consider the data presented in Fig. 3. In this experiment (Semb, Hopkins and Hursh,

Insert Fig. 3 about here

1973), the instructor measured student test performance on two types of items - study question items and non-study (probe) items. During the baseline condition, students earned points for correct answers to quiz questions which contributed to their grade in the course. Notice the advantage of short units - it enabled the instructor to obtain four data points before a second condition (Noncontingent Points) was introduced. During the Noncontingent Point condition, students were given the maximum number of quiz-points before they took the quiz. This was done to determine whether or not points awarded for correct responses promoted better performance. As Fig. 3 shows, there was a substantial decrease in mean student performance on both dependent variables. We could have stopped the experiment at this point, and

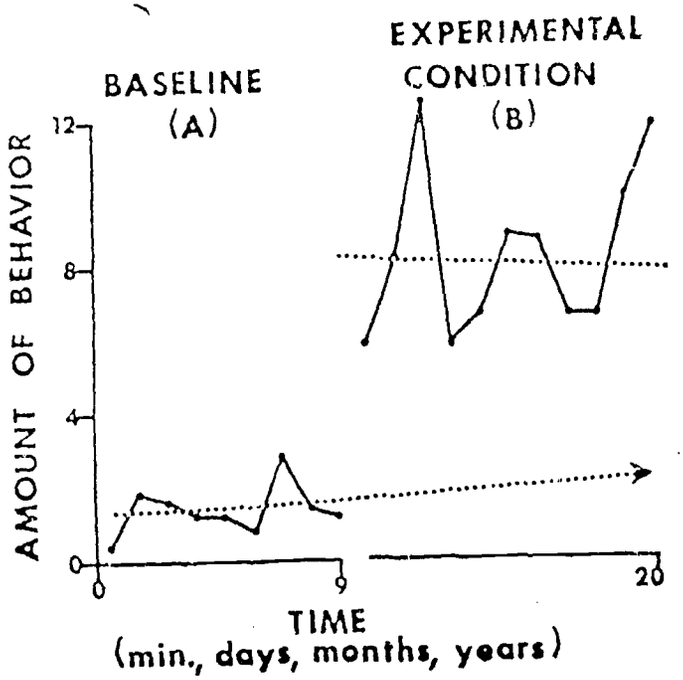

EXPERIMENTAL
CONDITION
(B)
BASELINE
(A)
AMOUNT OF BEHAVIOR
12
8
4
0
0
9
20
TIME
(min., days, months, years)

Fig. 1

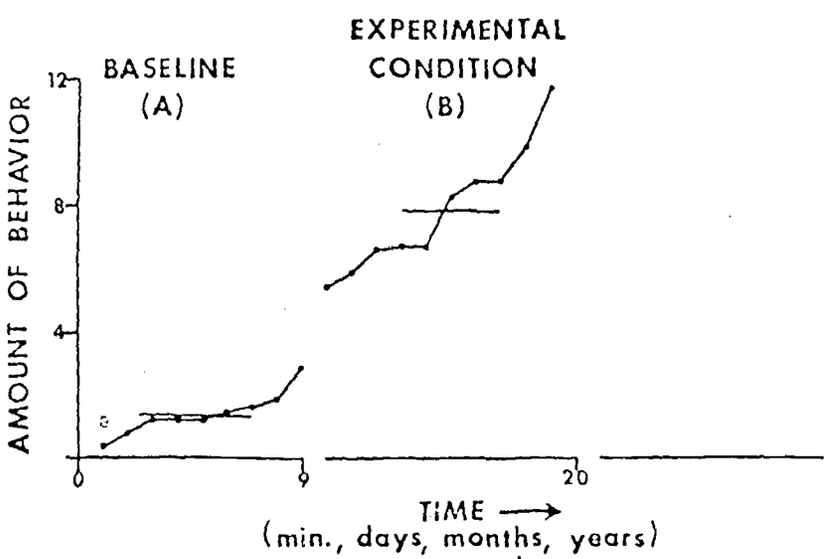

EXPERIMENTAL
CONDITION
(B)
BASELINE
(A)
AMOUNT OF BEHAVIOR
12
8
4
0
0
9
20
TIME ⟶
(min., days, months, years)

Fig. 2

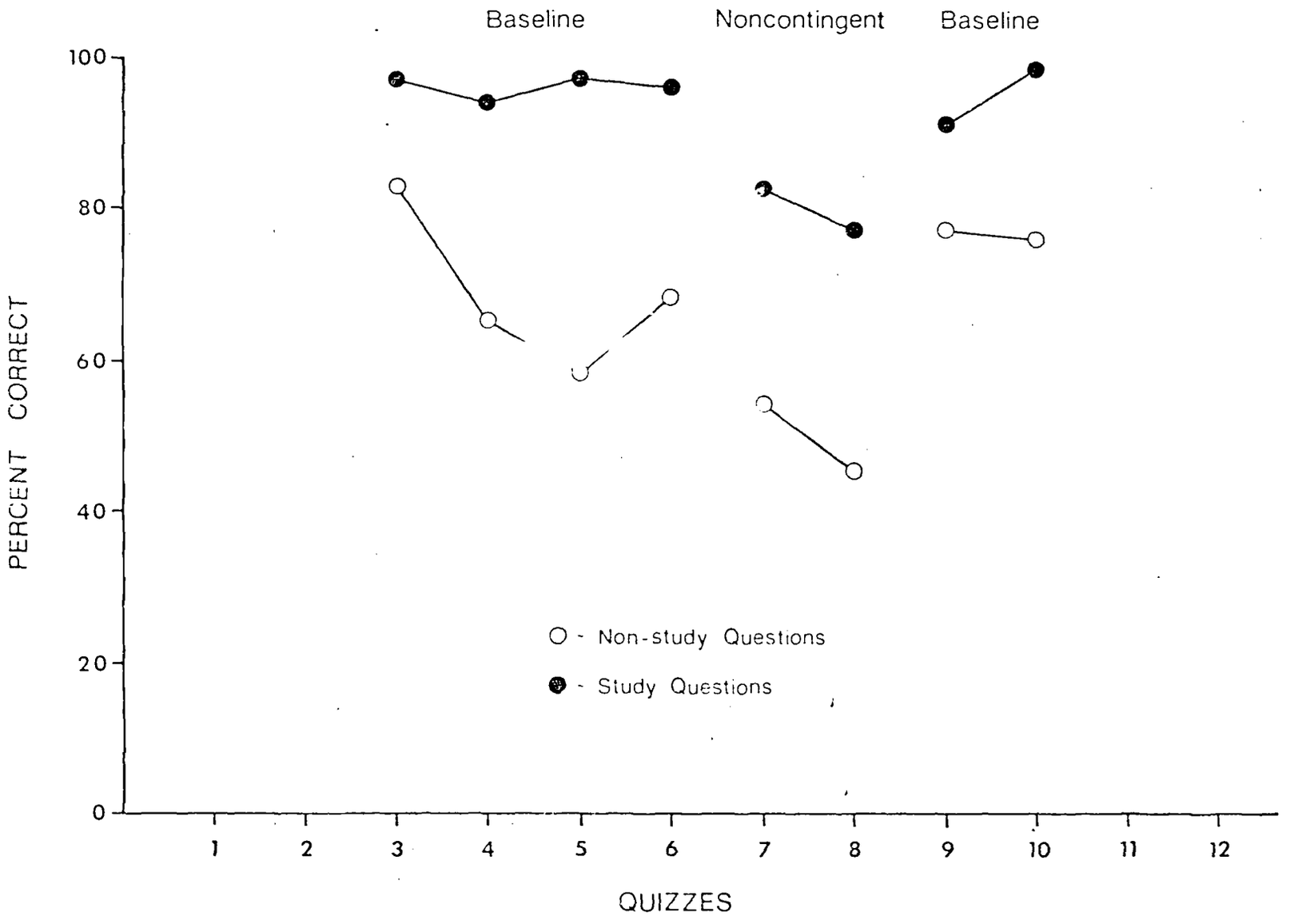

Baseline
Noncontingent
Baseline
100
80
60
40
20
0
PERCENT CORRECT
○ - Non-study Questions
● - Study Questions
1
2
3
4
5
6
7
8
9
10
11
12
QUIZZES

Fig. 3

we might have been relatively convinced that "free grades" lead to poorer test performance. However, in any time series analysis, it could have been due to a variety of other factors. For example, it could be that during the noncontingent point condition the material in Units 8-9 or Quizzes 8-9 were more difficult, or that students did not like the material, or that they were getting tired of the course. To increase our confidence that there was a functional relationship between points awarded for correct responses and performance, the original condition (Baseline) was reinstated, thus completing the ABA sequence. With repeated reversals we would add even more to the certainty that our manipulation actually produced or caused the observed changes in performance.

Notice, however, that the ABA reversal design, when applied to performance variables, suffers from a potential source of confounding. Because content and test items in each successive condition are different, it could be that changes in performance are due to changes in the difficulty of the content and/or test items. One way to control for this possible source of confounding is to "standardize" test items on another population of students who are exposed to the same treatment throughout the term (Semb, 1972; Semb, Hopkins and Hursh, 1973), or to use a panel of experts (e.g., graduate students). With a standardized pool of items, it is possible to select test questions with the same mean and range from condition to condition or unit to unit. Believe me, this is no easy task, and it involves an additional assumption - namely, that the difficulty of those items remains constant from one term to the next. Furthermore, caution must be taken to prevent items

from becoming part of student test files.

An alternative way to control for changes in item or content difficulty is to expose a second group of students to the treatments in opposite order from the first group. Bostow and Blumenfeld (1972) have used this design (see Fig. 4) and we make frequent use of it in

Insert Fig. 4 about here

our present experiments. We call it a counterbalanced reversal design. Both within and between group comparisons are possible. Because the two groups or subjects are in opposite treatments over the same content (counterbalanced), within group comparisons between successive treatments enable one to state that changes in performance or preference are a function of the treatment condition and not the content covered or test items used.

Another application of the reversal design is presented in Fig. 5.

Insert Fig. 5 about here

Miller, Weaver and Semb (1973) assigned target dates backed up by a course withdrawal contingency (pass the quiz by the target date or withdraw from the course) during the initial condition. Students averaged a little over one quiz completed per day. During the next condition, no target dates or contingencies were imposed on performance, and students' rates of lesson completion decreased considerably. To establish the functional relationship between target dates backed up by the course withdrawal contingency and student rates of lesson

EXPERIMENTAL DESIGN

ABAB
BABA

	Weeks 1 & 2	Weeks 3 & 4	Weeks 5 & 6	Weeks 7 & 8
DEFERRED POINT CONDITION	Group A	Group B	Group A	Group B
RAW POINT CONDITION	Group B	Group A	Group B	Group A

Fig. 4

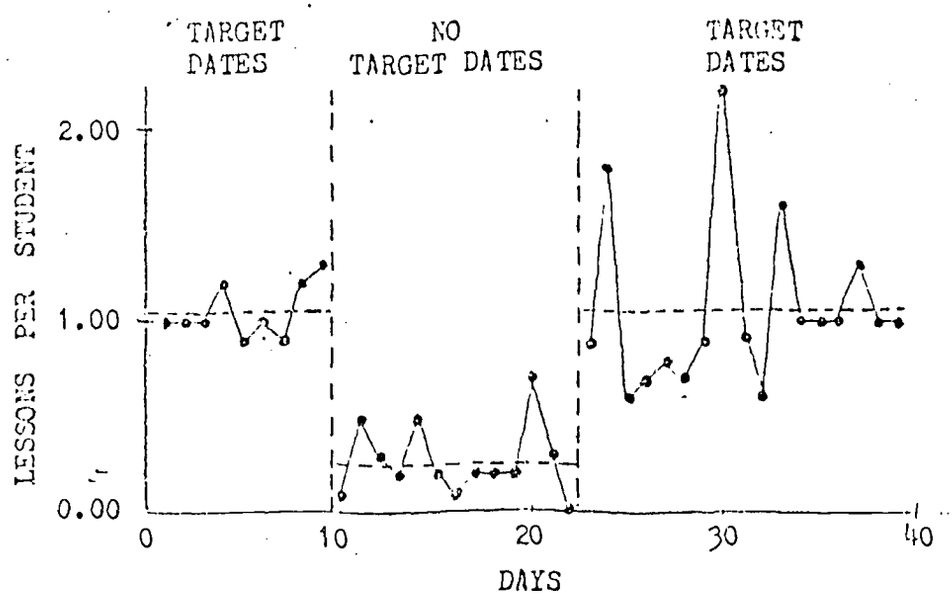

TARGET DATES
NO TARGET DATES
TARGET DATES
LESSONS PER STUDENT
2.00
1.00
0.00
0
10
20
30
40
DAYS

Fig. 5

completion, the initial condition was reinstated during the third phase of the course.

The reversal design is also nicely suited for measuring student preferences toward different teaching methods. It seems reasonable to suppose that if we want students to rate one method versus another that we expose him to both. This is certainly a major advantage of single subject experimental analysis, because in such an analysis the student is exposed at least once to both methods of instruction. If you are not satisfied with students' verbal reports such as course rating scales or questionnaires, there is yet another alternative with the reversal design. Everything remains the same as before in the first two (AB) treatments. However, during the final condition, the student is given a choice between A and B. This is an extremely strong measure of student preference in that regardless of what the student says, he or she must commit himself to one method or the other. Here we encounter the actuarial advantages of a large group of students - what proportion prefer each method? For an example of this research design, see Hursh, Wildgen, Minkin, Minkin, Sherman and Wolf (1973).

The reversal design is a powerful technique for analyzing individual student behavior. However, it has two major drawbacks. First, some behavior changes are not reversible. For example, once you have successfully taught a particular concept, there is no way to reverse the student's history. No number of reversals will produce the original level of the behavior. Second, in some cases, a reversal may be politically or ethically unwise. Alternatively, a manipulation may be so aversive that a return to the original condition occurs almost

immediately. For example, two years ago (Semb, 1972) I wanted to increase attendance at optional discussion classes following quiz days. Initially, we returned quizzes to students via a distribution center which was open all day. See Fig. 6 labeled baseline. To

Insert Fig. 6 about here

promote better attendance, we decided to return quizzes for the 1:30 lecture group only during discussion sections. As Fig. 6 shows, attendance increased. What it does not show is the wrath we encountered - mad, threatening students and an onslaught of phone calls. We had planned to implement the same procedure with the second group later in the semester, but scrapped the idea and returned to the initial condition for the 1:30 group. Sometimes, the situation is just the opposite. One starts with something bad, goes to something good, and then, for whatever reason, does not want to return to the bad. Fortunately, there is a design, the multiple baseline design, which can be used to establish causality over time without the necessity of a reversal.

Multiple Baseline Design

In a multiple baseline design, two or more behaviors of the same subject or group of subjects are measured at the same time. After initial baseline measures are obtained on both, the experimental manipulation is applied to only one behavior. The change obtained is compared with the level forecast for that behavior from its baseline. The confidence in our forecast is directly related to the level of the baseline of the second and subsequent behaviors being measured.

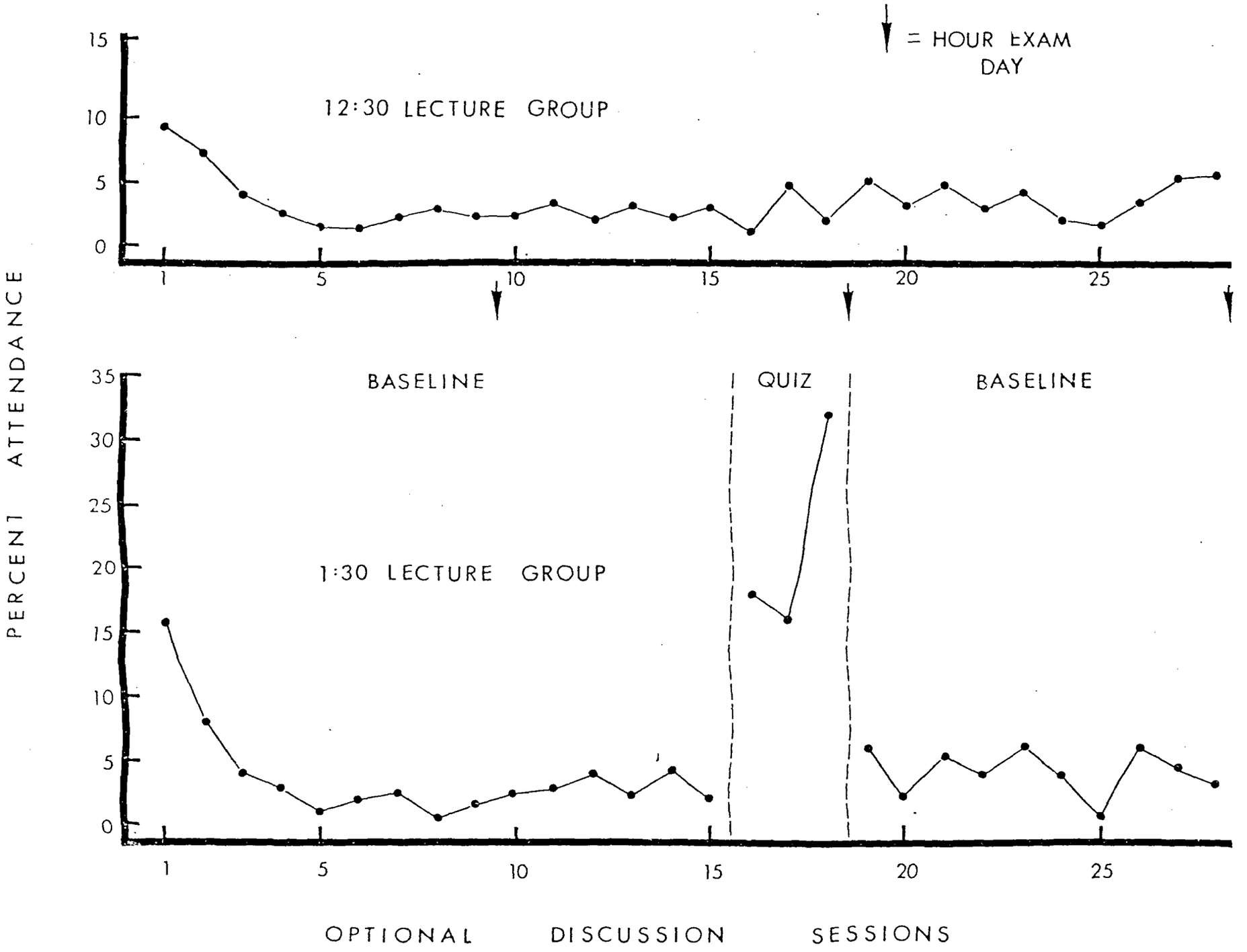

= HOUR EXAM DAY
12:30 LECTURE GROUP
1:30 LECTURE GROUP
BASELINE
QUIZ
BASELINE
PERCENT ATTENDANCE
OPTIONAL DISCUSSION SESSIONS
0
5
10
15
20
25
30
35
1

Fig. 6

Consider Fig. 7 (Miller and Weaver, 1972). Several times during

Insert Fig. 7 about here

the course (shown on the abscissa), Miller and Weaver administered the same, comprehensive, fill-in-the-blank achievement test. During the initial testing, students scored low on all five portions of the test (shown on separate graphs on the ordinate). As the course progressed, increases in performance were directly related to that part of the course which had just previously been trained. Notice, too, that because the test was comprehensive across the entire course, it allowed the researchers to obtain retention measures on material which had been presented early in the course.

The multiple baseline design is somewhat weaker than the reversal design in that it involves an additional assumption - namely, that all measured behaviors are susceptible to the same manipulations. This assumption can be supported, however, by applying the same procedure to many successive behaviors (Risley and Wolf, 1972). The comparison of interest in each is between baseline (pre-training) and treatment (post-training), not between sets of behaviors. The remaining behaviors serve to support the baseline forecast of the first behavior. Notice in Fig. 7 that the multiple baseline design allows the researcher to establish trends as well as mean effects. Performance on Conditioned Reinforcement material during baseline (pretraining) increased consistently, perhaps because the similarity between conditioned reinforcement and concepts presented earlier in the course was high.

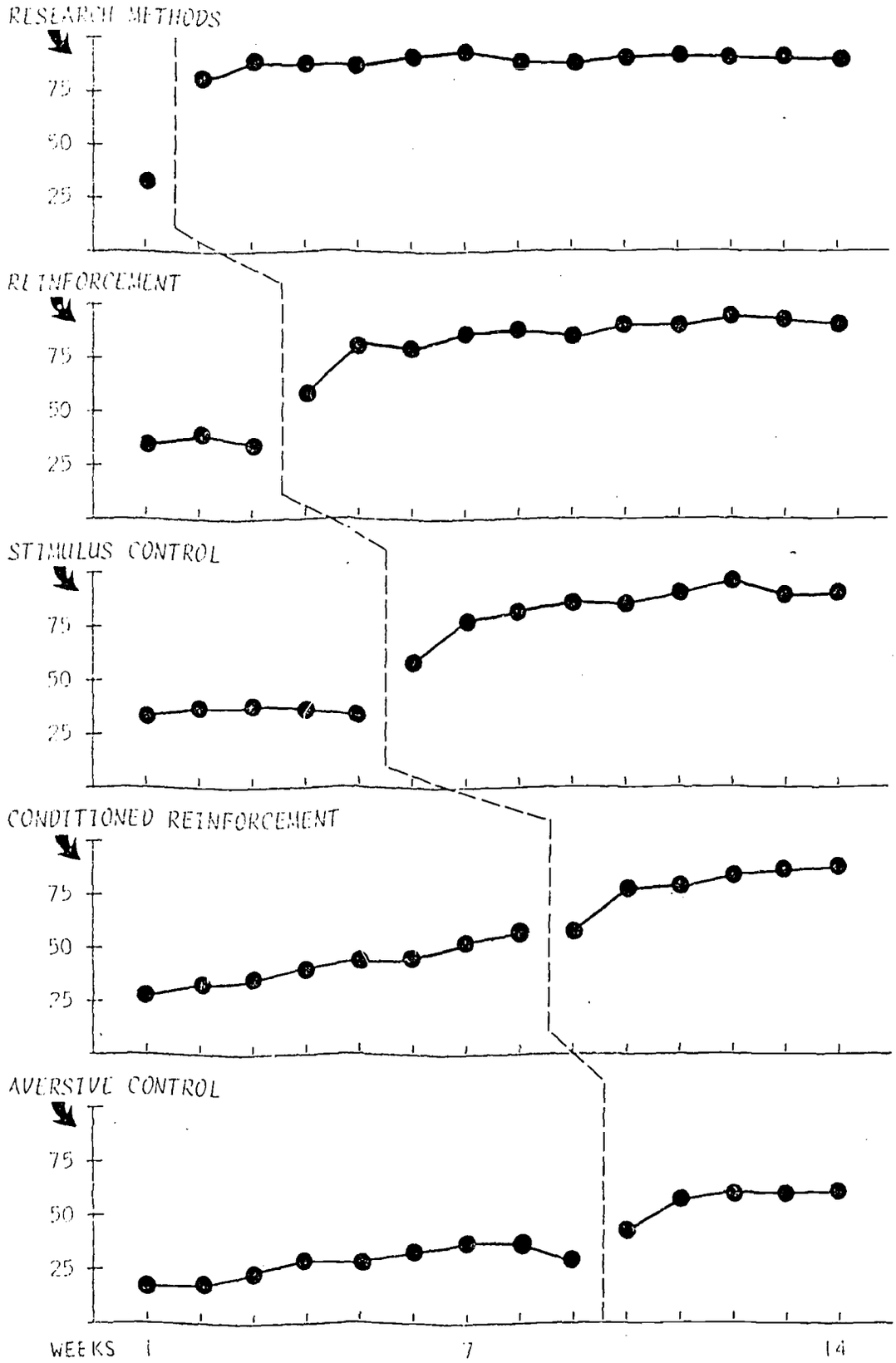

RESEARCH METHODS
75
50
25
REINFORCEMENT
75
50
25
STIMULUS CONTROL
75
50
25
CONDITIONED REINFORCEMENT
75
50
25
AVERSIVE CONTROL
75
50
25
WEEKS 1
7
14

Fig. 7

There are several types of multiple baseline designs. The two most useful variations in college teaching research include the one just described in which performance on all phases of course content are measured at frequent intervals with each part serving as a separate baseline. The second involves using two or more subjects or groups of subjects. The experimental treatment is introduced successively for each group at different times. For example, assume that each of the five coordinates shown in Fig. 7 represented a different group of students. At different points during the course, the experimental treatment is introduced first for the top group, then for the next group, and so on. Changes in behavior associated with each introduction of the independent variable increase our confidence that the manipulation produced the effect.

The Miller and Weaver (1972) multiple baseline achievement test can also be extended to compare the effectiveness of different teaching procedures. For example, in a recent experiment (Semb, 1973), I used the same type of comprehensive test administered five times during the semester. The data for two groups of students who went through the experimental treatments in different order are shown in Fig. 8. Although

---

Insert Fig. 8 about here

---

each treatment produced increases in performance on both study question items and non-study (probe) items, the magnitude of effect was not the same for each treatment. Furthermore, the graph is difficult to read, so we computed percentage gains in performance by subtracting pre-test scores from training scores. These data are presented in

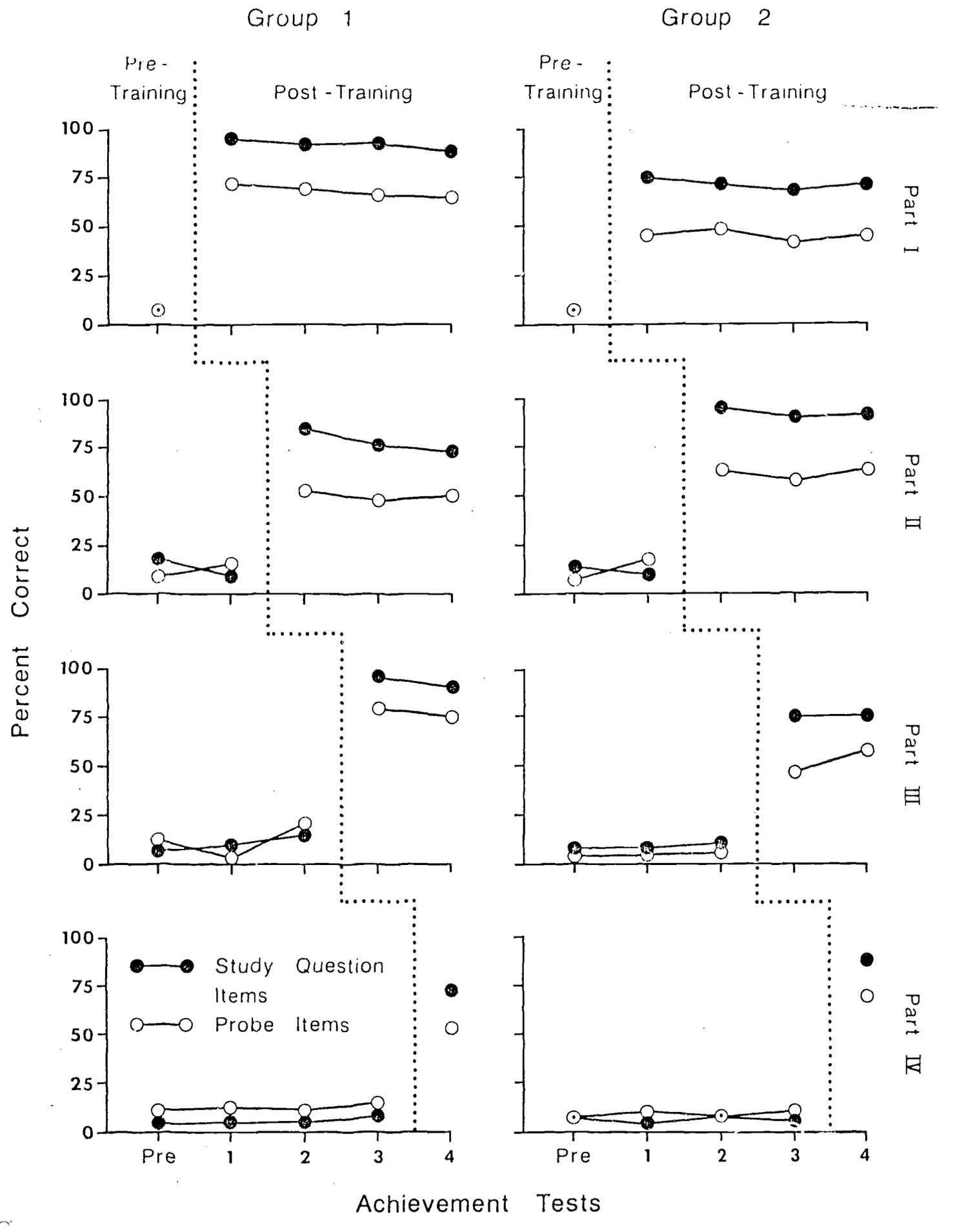
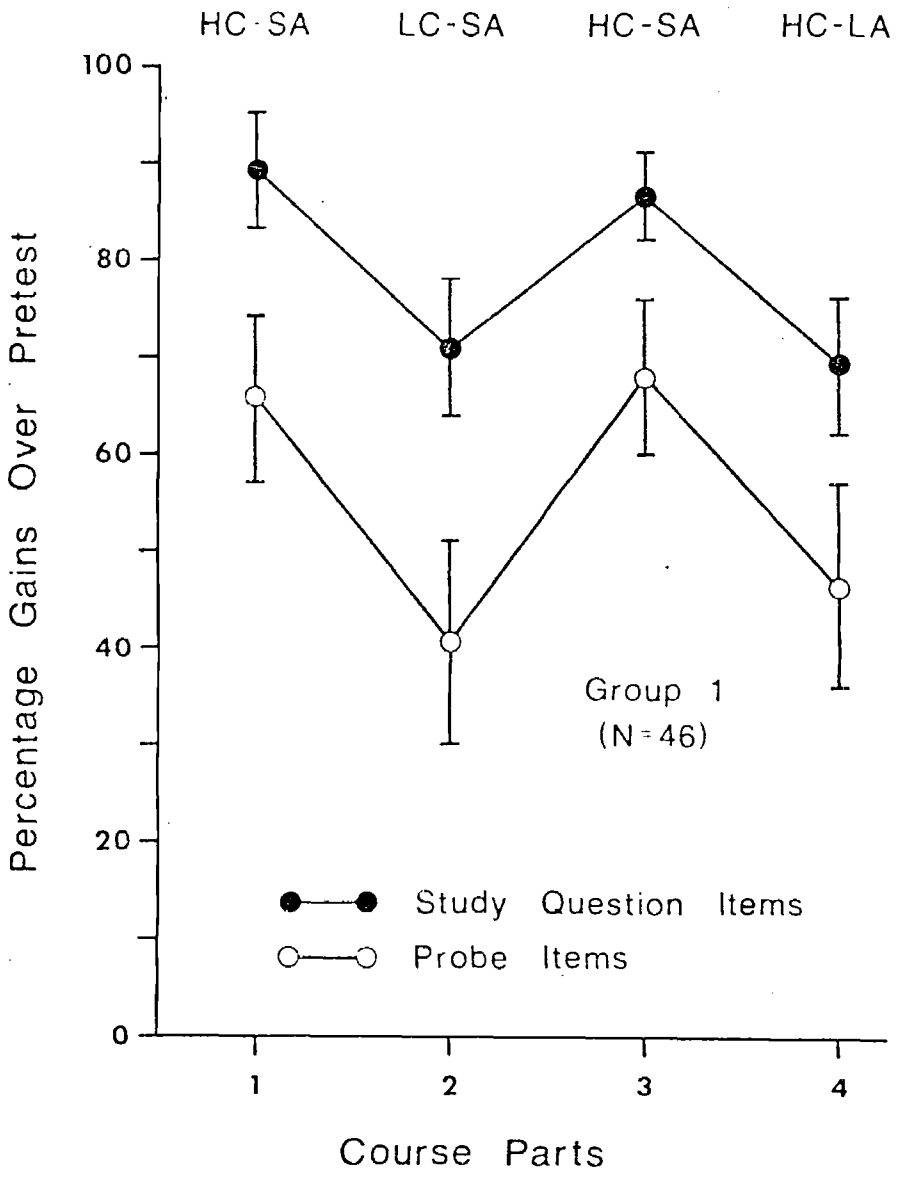

Group 1
Group 2
Pre - Training
Post - Training
Part I
Part II
Part III
Part IV
Percent Correct
100
75
50
25
0
Study Question Items
Probe Items
Pre
1
2
3
4
Achievement Tests

Fig. 9. Notice that the increases in performance over pre-test levels

Insert Fig. 9 about here

were greatest for the high grading criterion - short assignment condition on both types of items. Thus, it was possible, using the multiple baseline testing procedure, to demonstrate the functional relationship between each teaching package and performance, and furthermore, to compare different teaching methods as to their effectiveness.

Unfortunately, we have had to describe these experiments hastily, I would encourage those of you who are interested in using single subject analyses of the type described here to refer to the articles referenced for more detail. Let me reiterate my belief that single subject analysis has a great deal to offer researchers in higher education. Both reversal and multiple baseline designs involve applying two or more treatments to the same individual or group over time, thus enabling a precise analysis of individual performance and/or preference. Second, both allow the researcher to make statements about functional relationships between experimental treatments and observed changes in behavior. Finally, both have been demonstrated by the research results presented here to be viable tools for college teaching research.


HC-SA
LC-SA
HC-SA
HC-LA
Percentage Gains Over Pretest
100
80
60
40
20
0
Group 1
(N=46)
Study Question Items
Probe Items
1
2
3
4
Course Parts

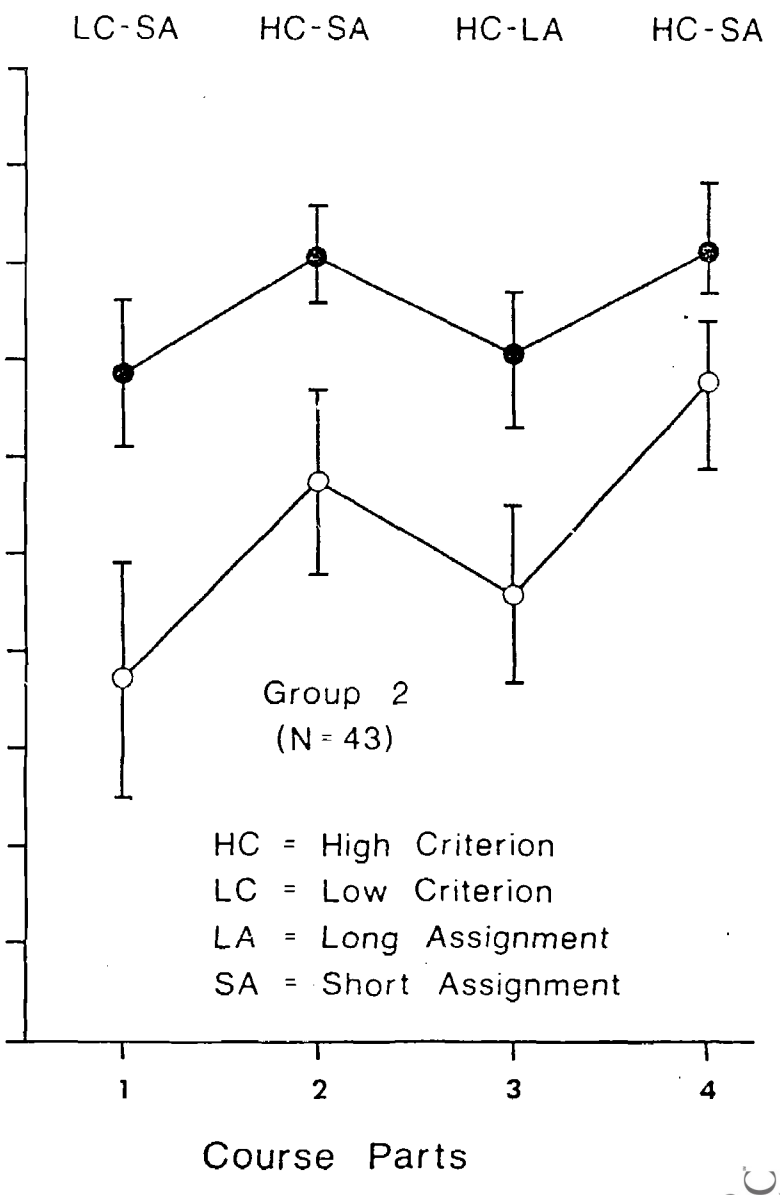

LC-SA
HC-SA
HC-LA
HC-SA
Group 2
(N=43)
HC = High Criterion
LC = Low Criterion
LA = Long Assignment
SA = Short Assignment
1
2
3
4
Course Parts

Fig. 9

References

Born, D. G., Gledhill, S. M., & Davis, M. L. Examination performance in lecture-discussion and personalized instruction courses. Journal of Applied Behavior Analysis, 1972, 5, 33-43.

Bostow, D. E. & Blumenfeld, G. J. The effects of two test-retest procedures on the classroom performance of undergraduate college students. In G. Semb (Ed.), Behavior Analysis and Education - 1972. Lawrence, Kansas: Support and Development Center for Follow Through, Department of Human Development, University of Kansas, 1972.

Campbell, D. T. & Stanley, J. C. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally and Company, 1963.

Dubin, R. & Taveggia, T. C. The Teaching-Learning Paradox. Eugene, Oregon: Center for the Advanced Study of Educational Administration, University of Oregon, 1968.

Hursh, D. E., Wildgen, J., Minkin, B., Minkin, N., Sherman, J. A., & Wolf, M. M. Proctors' behavior and students' performance in a self-paced (PSI) undergraduate course. Paper presented at the Second International Workshop on Behavior Modification, Lawrence, Kansas, February, 1973. Reprints: Dr. D. E. Hursh, Western Carolina Center, Enola Road, Morganton, N. C. 28655.

Miller, L. K. & Weaver, F. H. A multiple baseline achievement test. In G. Semb (Ed.), Behavior Analysis and Education - 1972. Lawrence, Kansas: Support and Development Center for Follow Through, Department of Human Development, University of Kansas, 1972. Pp. 393-399.

Miller, L. K., Weaver, F. H., & Semb, G. A procedure for maintaining student progress in a personalized university course. Journal of Applied Behavior Analysis, in press.

McMichael, J. S. & Corey, J. R. Contingency management in an introductory psychology course produces better learning. Journal of Applied Behavior Analysis, 1969, 2, 79-83.

Risley, T. R. & Wolf, M. M. Strategies for analyzing behavioral change over time. In J. Nesselroade and H. Reese (Eds.), Life-span Developmental Psychology: Methodological Issues. New York: Academic Press, 1972.

Semb, G. The effects of instructional objectives and grade-contingent points on student test performance in an introductory college course. Unpublished Doctoral Dissertation, University of Kansas, 1972.

Semb, G., Hopkins, B. L., & Hursh, D. E. The effects of study questions and grades on student test performance in a college course. Journal of Applied Behavior Analysis, 1973, 6, in press.

Sheppard, W. C. & MacDermot, H. G. Design and evaluation of a programmed course in introductory psychology. Journal of Applied Behavior Analysis, 1970, 3, 5-11.

Figure Captions

Fig. 1 - In the baseline condition the behavior of interest is first measured repeatedly over time. Then under the experimental condition the new level of behavior is measured repeatedly and compared with the level that would have been forecast from the baseline measure. (From Risley and Wolf, 1972)

Fig. 2 - The same mean effects presented in Fig. 1 now plotted in a manner such that the trends in the behavior under each condition are apparent. (From Risley and Wolf, 1972)

Fig. 3 - Mean percent of quiz items answered correctly for study and non-study question items during baseline and noncontingent points. (From Semb, Hopkins and Hursh, 1973)

Fig. 4 - An example of a counterbalanced reversal design. (From Bostow and Blumenfeld, 1972)

Fig. 5 - The mean number of lessons completed by each student during each day of the semester. The dashed lines represent mean lessons completed per day for each of the three experimental conditions. (From Miller, Weaver and Semb, 1973)

Fig. 6 - Mean percent daily attendance at optional discussion sections for the 12:30 and 1:30 lecture groups. The two experimental conditions included baseline and quiz return. Arrows indicate days on which hour exams were given in lecture. (From Semb, 1972)

Fig. 7 - Average scores on each subsection of the achievement test. Dotted line indicates introduction of teaching package for that subsection. (From Miller and Weaver, 1972)

Fig. 8 - Mean percent correct on study question items (closed circles) and probe item (open circles) on each of the five Achievement Tests. Group 1 is plotted on the left and Group 2 on the right. The four course parts are plotted vertically on separate ordinates and the five Achievement Tests are shown on the abscissa. (From Semb, 1973)

Fig. 9 - Mean percentage gains over pretest levels for each of the four parts of the course. Study items are represented by closed circles and probe items by open circles. Vertical lines through each point indicate standard deviations. (From Semb, 1973)

Footnotes

[1]Preparation of this paper was supported, in part, by grants from the University of Kansas Office of Research Administration (3317-5038 and 3316-5038) and the University of Kansas Office of Instructional Resources (71-2294). Reprints may be obtained from George Semb, Department of Human Development, University of Kansas, Lawrence, Kansas 66044.