

1975

## Frequent self-scheduled short quizzes in lieu of tri-semester mid-term exams.

Richard K. Wallin

*University of Massachusetts Amherst*

Follow this and additional works at: <https://scholarworks.umass.edu/theses>

---

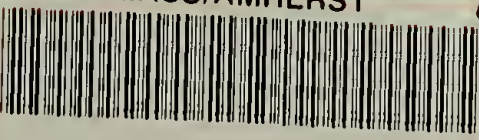
Wallin, Richard K., "Frequent self-scheduled short quizzes in lieu of tri-semester mid-term exams." (1975). *Masters Theses 1911 - February 2014*. 2058.

Retrieved from <https://scholarworks.umass.edu/theses/2058>

This thesis is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Masters Theses 1911 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact [scholarworks@library.umass.edu](mailto:scholarworks@library.umass.edu).



UMASS/AMHERST



312066013812109



FREQUENT SELF-SCHEDULED SHORT QUIZZES IN LIEU OF  
TRI-SEMESTER MID-TERM EXAMS

A Thesis

by

Richard K. Wallin

Submitted to the Graduate School of the  
University of Massachusetts in partial  
fulfillment of the requirements for the degree of

MASTER OF SCIENCE

March 1975

EDUCATIONAL PSYCHOLOGY

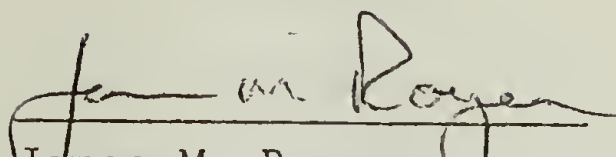
FREQUENT SELF-SCHEDULED SHORT QUIZZES IN LIEU  
OF TRI-SEMESTER MID-TERM EXAMS


A Thesis

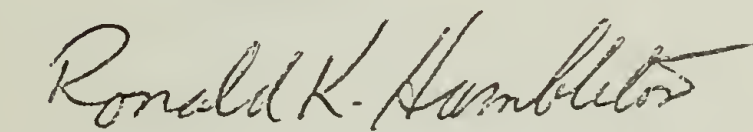
by

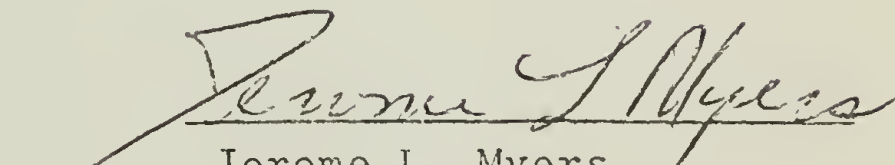
Richard K. Wallin

Approved as to style and content by:

  
James M. Royer  
Chairman of Committee

  
Harry Schumer

  
Ronald K. Hambleton

  
Jerome L. Myers  
Department Chairman  
Department of Psychology

## TABLE OF CONTENTS

INTRODUCTION	1
Two Successful Innovations	1
Research on the Effects of Frequent Testing	4
Results Not Showing an Advantage To Frequent Tests	8
Self Pacing	13
Rationale of the Study	14
Interaction Effects	15
METHOD	16
Subjects	16
Materials	16
Design	18
Procedure	19
RESULTS	21
Aptitude Treatment Interactions	31
DISCUSSION	40
APPENDIX 1.	46
REFERENCES	47

## INTRODUCTION

Too often in higher education, innovations are adopted or abandoned with little regard for documentation of their effects. Mindful of this problem, this project was designed to study a particular innovation planned to improve the efficacy of a typical college lecture course, one where the student is required to learn particular facts and concepts either straight from the lecturer or from certain detailed and carefully specified supplementary readings. This introduction will provide brief descriptions of two well-known and successful innovations out of which this project grew, will outline reasons why the innovation described herein might be of more practical importance in certain circumstances than either of the more well-known innovations, and an attempt will be made to detail some of the data which seemed to indicate that this innovation might indeed be facilitative of student learning. There will also be a discussion of the main features of the experimental design, and of some supplementary personality traits which were examined for possible interaction effects with the main treatment.

### Two Successful Innovations

One of the most successful recent innovations in higher education has been the Keller method. In this method (Keller, 1968) the instructor divides his course into fifteen or more distinct units of textual material and then requires that each of his students be tested on, and show mastery of, each unit, before he is allowed to proceed on

to the next unit. This requirement is not nearly as anxiety provoking as it may sound, as each student is given as many opportunities as he needs to show mastery, there being many more than one form of each test, and there being no limit to how many times he can take them. These tests are graded by course assistants, generally undergraduates, immediately upon their completion, with the student sitting next to the assistant during the grading, and encouraged to discuss his results with the assistant thereafter, in order to clear up any confusion he may have had with the material. This method, and many variations of it, have been shown to be able to increase student achievement in a course by a significant amount, it not being uncommon for a course of this type to have more than 50% of the students receive a criterion-referenced grade of A (Kulik, Kulik, and Carmichael, 1974).

Another successful approach to improving college teaching has been the mastery approach used by Bloom (1968). In this approach, students are expected to take 'formative evaluation' tests at intervals throughout the course. These tests are designed to assess the degree to which the student has mastered the material in the learning unit covered by the test. For those students who have thoroughly mastered the unit, the formative tests should reinforce the learning and assure the student that his present mode of learning and approach to study is adequate. For students who lack mastery of a particular unit, the formative tests should reveal the particular points of difficulty. The teacher should then, on the basis of this diagnosis, refer the student to particular instructional materials or processes intended to help



him correct his difficulties. These formative tests also provide useful feedback to the teacher since they can be used to identify particular points in the instructional process that are in need of modification.

Using this technique, Bloom improved student achievement in a test theory course quite significantly. In 1965, before using this technique, 20% of his students had achieved at an "A" level. In 1966, the first year he used his technique, 80% of his students received an "A" grade on a parallel exam. In 1967, this percentage increased to 90%.

While the effectiveness of both these techniques is indisputable, they both suffer from one major failing. In order to use them, the teacher is required to overhaul his course in rather significant ways, such as abandoning lectures, articulating instructional objectives, using student proctors, and so on. For a professor whose main interests lie elsewhere, this overhaul is often more extensive than he is likely to want to undertake. Thus, one of the goals of this project was to come up with an innovation that did not require such a massive change in the basic fabric of a course. In fact, the innovation described here is one that could easily be affected by a professors' assistants, since it requires no actual change in the way material is presented to, or discussed by, the class.

The key to such an unobtrusive manipulation seemed to lie in varying the conditions and scheduling of the mid-term exams in the course. Besides the obvious centrality of frequent testing to both the Keller and Bloom techniques, there is a supply of data that seems to show that frequent testing, by its very existence, facilitates



learning. In the following section, the reader will find a critical review of this research on frequent testing.

#### Research on the Effects of Frequent Testing

The research on the effects of frequent testing is noted for its variety of style and design. One of the more positive studies was done by Fitch, Drucker, and Norton (1951) who ran a study using two sections of an advanced course in government at Purdue University. Their control section, of 97 students, was given regular monthly exams in addition to their three lectures a week. The experimental section, of 186 students, was given a ten minute objective quiz at the end of the third lecture each week, in addition to the monthly exams given to the control group. In addition, both groups were allowed to attend a weekly, optional, discussion group. Fitch et. al. used, as their criterion measure for success in the course, the cumulative grades from the four mid-terms and the final. The second and third mid-terms were essay exams, while the first, fourth, and final exams were objective item tests. All students were given grades corresponding to their positions on a normal curve. Their final mean scores were 62.7 and 55.5 for the experimental and control groups respectively. It was noted that attendance at discussion groups correlated with both high grades, and frequent examining, but even after this effect had been partialled out, there was still significant advantage noted for frequent testing.

On a smaller scale, Turney (1931) did a study comparing two sections of an educational psychology course, each having around 40

students. On the first day he administered one form of the final for the course, consisting of 90 true-false items, 10 multiple-choice items, and 75 points worth of one word and completion items. One section scored 20% lower on this test and was consequently chosen to be the experimental group. (Note that no difference was detected between the sections on a mental aptitude test.) Both groups were given a 164 point mid-term exam, and a final (the criterion measure) which consisted of both forms of the final that had been created, that is, both the pre-test, and a parallel form that was new to the students. The control group was also given one other short exam.

The experimental group was given 12 short quizzes during the semester, 11 of which had items suitable for points scoring. These tests were not returned to the students. As no practice effect was detected on the first half of the final, it was included as half of the criterion measure. On this criterion measure no differences were seen between the two groups. However, because the two groups had started off at different levels of knowledge, as measured by the pre-test, Turney maintains that the experimental group benefitted by the treatment, as is shown by their 16% higher gain score. Unfortunately, this conclusion is open to debate, as it is not clear that the difference between the two groups on the pre-test can be reliably attributed to a difference in original knowledge between the groups.

Turney was not the only researcher who used gain scores as his criterion for success. Kulp (1933) ran a graduate course in educational psychology in which his 32 students were all given weekly exams on their

material. They were then given a mid-term exam which had on it the same questions as were on the small quizzes. Those students who scored in the top half of the class were subsequently excused from the requirement of taking the weekly quizzes for the rest of the semester. On the final at the end of the semester, there was no significant difference between these two halves of the class. Kulp states that the bottom half of the class gained ten points over their earlier average, while the top half lost ten points, thus obliterating the 20 point differential that had existed as mid-term time. Two problems with this research were, first, that no evidence was presented to explain why the two tests should be considered to be parallel, and second, that, according to Keyes (1934), most of the loss of the difference can be accounted for by regression to the mean.

On occasion, added innovations can confound any conclusions one might draw from an experiment on frequent testing. Smeltzer (1931) had just such a problem with an experiment he did on a large class in educational psychology, a class divided up into a few sections. In this experiment, the experimental section was given a 20 minute objective test every Thursday. This test was graded on Friday and those who had scored an A or B were excused from attendance at Monday's discussion. On Monday, the test was reviewed, and another 20 minute objective test was given as a retest. Each student's grade was the average of these two tests. On the final the median scores of the two groups were 230.6 and 222.0 for the experimental and control groups respectively. While this increase held for the entire class, it was



interesting to note that there was a marked advantage to the worse scoring students in the experimental group. The scores for the bottom tenth percentil were 202 and 172 for the experimental and control groups respectively. That this was so seems quite reasonable when one considers that these students received the bulk of the advantage from retaking the exams on Monday. Unfortunately, that these students derived the greatest benefit also puts in doubt the conclusion that the experimental group's advantage in the course was due to the frequent testing. More likely, it was the result of the detailed review sessions held every Monday.

While studying the Keller method, Mertens (1971) did an experiment which incidentally studied the effect of frequent testing. He took four very large sections of introductory psychology, and put them into four different treatments. These treatments were: 1) lecture but no text; 2) lectures and test; 3) daily testing, mastery not required; and 4) daily testing, mastery required. The medians on the final for these groups were 29, 59, 64, and 95, respectively. While the difference between the medians of the two groups that concern us here, groups 2 and 3, is not very large, it is statistically significant. Also, Mertens does not make clear whether the grades on his test counted at all, if mastery was not required. It may be that he was merely controlling for exposure and that these tests were not considered important by the students.

There are a number of reasons why frequent testing may, in fact, be facilitative of student learning. It may be that frequent testing

conforms to the operant model suggested by Bloom where a good result reinforces good study strategies, a sort of 'learning to learn' effect. It also might be the natural result of a quite simplistic model of student study behaviour. This model hypothesizes that the student is likely to leave all of his work until the last possible moment. Assuming that these 'last moments' have a finite durational limit, it seems logical that a small segment of work will capture a proportionately larger amount of study time. Thus, frequent testing, a device which breaks the semester into smaller segments for study purposes, is likely to encourage a larger total amount of study time put in by each student.

#### Results Not Showing an Advantage to Frequent Testing

A few of the studies about frequent testing do not support the hypothesis that more frequent testing causes greater course achievement. Wiggins, Pope, and Bushell Jr. (1968) did a very complicated study during which they examined not only frequency of testing, but also the weighting of quizzes, rewards for performance (such as movies, being in an honours class, excused class attendance, and exemption from quizzes and finals), scheduling of quizzes, and whether or not quizzes were preannounced. In this study they used six sections of a course in learning (mostly behaviour modification) taught by more than one instructor. Because of the difficulty of controlling for instructor style, they used an ABA design. In such a design the experiment is divided up into three time periods, and each group is exposed to more than one treatment. In this way each group serves, for statistical purposes,

as its own control. This strategy was a real weakness, since none of their groups were under frequent testing conditions for more than five weeks at a time. These five weeks corresponded to the time between major mid-term exams. Also, the conditions in this experiment did not differ greatly in the degree to which they exposed the students to frequent testing. Given three five week periods for each of six sections there were, according to their calculations, eighteen conditions.. Of these, two included no tests, nine included two tests, and seven included four tests, hardly a substantial variation across conditions.

In this study there were five different measures of achievement. There were two objective mid-term exams of 40 items each, and a 30 item objective final. They also gave a 20 item pre-test to each student, and retested using this same test just before the end of the course, without preannouncement. In addition, a 35 item test was given to a sample of the upper quartile of the course fifteen weeks after its completion. Analysing their data in a singularly complicated manner, they reported eighteen comparisons between frequent and non-frequent testing, of which ten came out in the expected direction, seven came out showing no difference, and one came out in the wrong direction. According to their calculations, this data was not consistent enough to attain any sort of statistical significance. As an added fact, they reported that there was not a significant difference between pre-test scores and scores on the test given fifteen weeks after the end of the course, a result hardly encouraging to any educator.

A more simple and straightforward study was run by Hertzberg,



Herlmann, and Leuenberger (1931) who compared the results they got with their educational psychology class in the fall (the control) with results they got using frequent testing in their spring semester class. During the third class every week in this experimental section, they gave a short quiz, consisting of true-false, multiple-choice, and completion items. They then marked these quizzes and returned them so that the students could use them as study aids. The students were examined three times, being given two mid-term exams and a final. On the two mid-terms the experimental group did 15% and 12% better. However, on the part of the final that concerned these first two thirds of the course there was no difference in scores from one semester to the next. This may be the result of the fact that at each mid-term the instructors collected the quizzes for that period of the course, so that the students did not have access to the quizzes as study aids for the final. It should be noted that, in addition to the differences in mean scores on the first two mid-terms, the experimental group also had a smaller standard deviation of scores, it being 75% as large as the control's on the first exam, and 80% as large on the second. This would seem to indicate that the frequent testing was even more helpful to the poorer students than it was to the better students.

Bostow, Mawhinney, Laws, and Blumenfield (1970) describe two experiments they conducted to determine the effect of frequent testing on study behaviour. In both experiments they took a few students (eight and twelve, respectively) from a course in educational psychology and had them do all their studying in a room equipped with an observation

window. They were not allowed to take any of their books or notes home with them, but were allowed to study from them in this special room for as much time as they wished during the hours of 3 to 6 on Monday to Thursday. In the first experiment they were given daily quizzes during weeks 1,2,6, and 9, and weekly quizzes at the end of weeks 4,5,7, and 8. In the second experiment, they were given daily quizzes during weeks 1,2,6, and 7, and tri-weekly quizzes at the end of weeks 5 and 10. They found that the students subjected to daily quizzes studied more consistently than those on less stringent schedules. They also found that over the longer intertesting periods students tended to leave their work until the end. While this result confirms the premise stated earlier that students tend to leave their work till the end, Bostow et. al. neglected to sum the studying that their students did, so it is impossible to say whether they ended up doing more or less in total as a result of frequent testing. They reported no differences in achievement between any of the modes.

A study which closely resembles the one described here was done by Keyes (1934). Keyes noted that much of the research done on frequency of testing does not really control for exposure to the material. He noted that in most of these studies the students receiving frequent testing were also being exposed to more items, were being given more review sessions, and were demanding more teacher attention. To remedy these problems, he divided his educational psychology class of 286 students into two groups, carefully matched for sex and score on a 167 item true-false pre-test. He then gave his experimental section a weekly test on

that portion of the material. After five weeks he gave his control group a mid-term exam consisting of all the items that had been given to the experimental group, thus controlling the number of items each group saw. He did this for the first two five week periods of the semester. During the third five week period he gave only an end test to both groups. It should be noted that all of these mid-semester tests contained both true-false and completion items in a ration of 7:1. Two weeks before the end of the course he administered a surprise test consisting of the 118 items from the pre-test that were covered in the first two thirds of the course. Then during the finals period he administered a true-false final to both groups.

The experimental group did 12% better on both of the first two periods. However, they also did 6% better on the test at the end of the third period. Keyes hypothesized that some of their better study habits may have stayed with them. On the surprise post-test the experimental group scored 7% higher than the control, but on the final exam there were no differences between the groups.

Keyes also took an attitude survey at the beginning and end of his course. At the beginning he found that 45% of the two groups wanted frequent (every 2,3,or 4 days) testing, and 37% were happy with only monthly tests. By the end of the semester the number wanting frequent testing had increased to 59% while the number favouring monthly tests had decreased to 24%.

One common characteristic of all these experiments is the lack of consistency throughout the whole semester. None of the researchers



maintained their experimental procedure for the whole semester, and it is felt that this is likely to be the reason that none of them got positive results when using a final exam as their criterion.

### Self-Pacing

A second feature of the design used in this project was that it allowed students to set, to a certain degree, their own pace of exam taking. There is some reason to believe that an amount of self-pacing of exams in a course is facilitative of learning. Keller (1968) insists that self-pacing is one of the features of his method that makes it as successful as it is.

Born (1970) did an interesting experiment which, though it showed a positive trend, failed to support the idea of self-pacing. Instead of running a normal Keller method course in his introductory psychology course (as in Keller, 1968) he divided the textual material into 57 units, and allowed his students to be examined on as many of these units as they wished to be at one time.

Born also reported that his students had a good attitude towards the self-pacing component of his course. This author believed that such a good attitude would be likely to translate into higher achievement, and, for that reason, if none other, felt that self-pacing was worth incorporating into the design.

## Rationale of the Study

The purpose of this study was to provide documentation on the effects of the educational innovation described in this paper. The innovation was studied to determine whether it had any effect on the amount of course material the students learned or whether it influenced student attitude towards the course.

The innovation (described in more detail later in the paper) was basically a manipulation of the conditions and scheduling of the mid-term exams in the course. Instead of having a fixed schedule of three exams, one each month, the experimental group had the option of taking their exams in smaller pieces, and had a choice of four different dates on which to take the test for any particular piece. In the extreme, a student might well have opted, as some did, to take twelve weekly tests covering, in sum, the same material as was covered in three tests taken by the control group. On the other hand, he might have decided to limit himself to the minimum of three exams, the traditional pace.

There were a number of advantages to this innovation over and above the ones of frequent testing and self-pacing. First, a student was usually not obliged to take an exam on a very inconvenient testing day. Were he to have another exam or a pressing engagement coming up, he would be able to plan ahead and get his examining done for this course during a more convenient week.

A second advantage of this particular design is that although it seemed to, and did, give a student a wide option in the number of testing days he attended, the structure of the innovation was such as to

encourage the student to come more often rather than less often. If one subscribes to the notion that students leave all of their work until the very last moment, it is easy to see that, starting with the fourth week, there would have been a 'last moment' for one week's worth of work each week. Unfortunately, this effect seemed not to be important in the actual running of the experiment.

### Interaction Effects

There were two personality traits that were examined for possible interaction effects with the innovation. The first was test anxiety, that is, the amount of anxiety a person displays when taking a test or even just thinking about it. Intuitively, it seemed that test anxiety could have affected the results of this experiment in either of two directions. It might have been that the smaller tests would cause less anxiety for students normally anxious about tests, or, conversely, it might have been that students high in test anxiety would limit themselves to the minimum number of testing situations, thereby negating the positive effects of frequent testing.

The other trait examined was 'internal vs. external control'. This trait measures the extent to which a person feels that his actions can control the important outcomes in his life, in this case the grades in the course. It seemed reasonable to assume that a person who measured high on this trait would respond favourably to the options presented to the experimental group. Also, Wiggins et. al. (1968) reported that people who scored very internal on their scale seemed to study more under conditions of frequent testing even though they, seemed to do no better on an objective final.



## METHOD

## Subjects

The subjects in this experiment were the students enrolled in an adolescent psychology course at the University of Massachusetts. Although there were between 450 and 500 students in the course, complete records were only available for 394. All the students were initially randomly divided into either the control or experimental groups on the basis of the last number of their student numbers. Some students had classes that conflicted with the testing time for the experimental group and these students were added to the control group. The final numbers of students were: real control group - 199; real experimental group - 140; experimental group people who, because of a conflict, were added to the control group - 55.

## Materials

The major materials in this experiment were the test forms. The semester's work (lectures and 50 assigned readings) were divided into twelve equal segments, each corresponding to roughly one week's worth of material. Four test forms were made for each of these twelve segments, each having twelve questions probing roughly 7:5, reading versus lecture material. The questions were drawn mostly from a pool of items developed over four semesters of teaching the course. They were randomly distributed among the test forms with the one proviso that each form should have at least one question from each of the readings in that segment.

Some of the questions had been item analyzed in previous semesters, and the forms were checked to have comparable difficulty by comparing these questions.

The mid-term exams given to the control group were a compilation of the four forms given to the experimental group on weeks four, eight, and twelve respectively (see figure 5). To make this a little clearer, let us take an example of week eight. The mid-term exam given to the control group on the test day of that week consisted of the fourth form of the tests on the material covered in the fifth segment, the third form of the test on the sixth segment, the second form of the test on the seventh segment, and the first form of the test on the eighth segment.

The main dependent measure, the optional final exam, was a 25 item test covering material presented in the first six segments of the course. These items were drawn from the pool earlier in the semester so as to be a representative sample of the ones from which the quizzes were drawn. There were at least two items included in that test that pertained to each segment.

The item selection was limited to these first six segments for two reasons. First, this helped keep the two groups, experimental and control, equivalent in their recency of exposure to the material. Second, by testing only the first half of the course material, it was possible to view the experimental test as a measure of retention, a measure considered more meaningful than normal final exam score.

Two of the personality trait tests were short forms developed by Wiggins et. al. (1968). Their test anxiety scale was a shortened version

of the Alpert-Haber Test Anxiety Scale. Their internal versus external control scale was a shortened version of the Rotter-Seaman-Liverant Internal vs. External Control Scale. Both short forms had three Likert type questions with a five point scale. Both tests were considered by Wiggins et. al. to be quite discriminating and reliable.

The other anxiety scale used was the Anxiety Differential developed by Alexander and Husek (1963). This was an eighteen item semantic differential test developed especially for determining test anxiety.

The course evaluation instrument was one developed recently at UMass to be used to evaluate all courses at the university. It consisted of twelve main items, and a few subsidiary background items. (See Appendix 1.)

The form that contained both this optional final and the personality trait tests also questioned the students as to how much time per week they put into studying for the course, and various other demographic data that were considered by the author to be reasonable targets of opportunity.

## Design

The experimental design was rather simple. No pre-tests were given. There was a control group and an experimental group and each was kept in its respective condition throughout the whole course. There were two post-tests, the optional final, and the student evaluation. These were given at the identical time and in the identical form to a large, random sample of both control and experimental subjects.



The optional final was intended to probe retention of course material. It was given on the last day of class, at least a month after the average person in both experimental and control groups had been tested on the material covered by the test. It was hoped that the subjects who took this optional final would be a representative sample of the whole class. In fact, more than three-quarters of the class did attend the optional final.

### Procedure

On the first day of the course the design of the experiment was explained to the class. There seemed to be no adverse response. A sheet detailing the arrangements was handed out to all students.

On each Wednesday night of the semester, except for the first, there was a testing session in a large lecture hall. At this session tests were administered to those wishing to take them, providing they were in the experimental group. They could take all four of the tests available on that day, or they could take three, or two, or one, or, of course, if they didn't come, none. The only limitation on missing testing days was that the student was required to take at least one of the four forms of each test so that it was impossible to skip more than three testing days in a row. For instance, let us say a student missed the first testing day but took the first test on day two. He then missed the third day and took the second test on testing day four. He then came in on the fifth testing day and took tests three, four, and five. Seeing as he was caught up, he could then skip three testing days, but would

have to come in, at the latest, by the ninth testing day to take at least the test on the sixth segment. It proceeded in this way for the whole semester.

On the testing day of the fourth week and of the eighth week all students in the control group took their normal mid-terms. The control group took their final exam during finals period, and the experimental group was offered one last testing session at the same time.

The optional final was administered to the whole class, at least those who came, on the last lecture period of the semester. Each student who came was offered a small number of extra points towards his final grade in the course. This offer was not made conditional on performance on the test. The students were urged to try their hardest on the test, and the interest displayed in the return of the scores, and the scores themselves, indicate that the students did, in fact, try hard on the test.

All of the tests, the ones given to the experimental group, the ones given to the control group, and the optional final, were scored by computer and the results posted prominently within one or two days of the testing day.

## RESULTS

A 't'-test between the means of the two groups on the 25 item optional final (14.66 and 15.21 for control and experimental groups, respectively) yielded a result which indicated that there was no significant difference between the two means ('t'=1.26,  $p=.10$ ; see Table 1.)

Inspection of the data revealed that there was some bias in the result reported above. As has been explained, each student, whether in the control or experimental group, took twelve segment tests of twelve questions each. Therefore, for their total score in the course, each student had a possibility of 144. While in the entire class the experimental group ( $n=140$ ) had a higher average total score in the course than the control group ( $n=254$ ) (means are 106.1 versus 105.6), in the sample of those who took the optional final this ranking was reversed (control-107.1, experimental-105.8). Because score on the optional final correlated very highly with total score in the course (control:  $n=186$ ,  $\text{corr}=.71$ ,  $p<.001$ ; experimental:  $n=116$ ,  $\text{corr}=.66$ ,  $p<.001$ ) it was decided to do a one-way analysis of covariance comparing the groups on their optional final scores with their course total score covaried out. This done, a significant difference favouring the experimental group was established at the  $p<.02$  level ( $F=5.678$ ,  $df=1/299$ ; see Table 2).

It was hypothesized that the experimental testing procedure might affect the students' attitude towards the course in general as measured by the UMass Provost Evaluation Form. Because it was assumed that the



TABLE 1.

## DIFFERENCES BETWEEN GROUPS ON MAJOR MEASURES

MEASURE NAME	MEANS (s.d.'s in parens)		't'	p < x
	CONTROL (n=186)	EXPER. (n=117)		
Score on Optional Final	14.66 (3.75)	15.21 (3.44)	1.26	.10 <sub>a</sub>
Total Score in Course	107.1 (16.2)	105.8 (15.2)	.695	.25 <sub>a</sub>
SAT-Verbal (self-report)	551.8 (76.3)	548.9 (72.1)	.333	.65 <sub>b</sub>
Grade Point Average	2.99 (.42)	2.97 (.43)	.419	.62 <sub>b</sub>
Work (hours per week)	2.58 (1.8)	5.08 (2.2)	10.7	.001 <sub>b</sub>
Anxiety Differential (high no.= high anxiety)	26.0 (8.9)	26.4 (9.6)	.37	.52 <sub>b</sub>
Anxiety Short Form (high no.=low anxiety)	5.65 (2.9)	5.48 (3.0)	.494	.44 <sub>b</sub>
Internal vs. External Control (high no.=internal)	6.42 (2.7)	6.93 (2.9)	.872	.40 <sub>b</sub>
Total No. of Tests Taken	3	7.97 (2.13)	-----	-----

TABLE 2.

## ANALYSIS OF COVARIANCE TABLE

SOURCE	DF	YY	SUM-SQUARES (DUE)	SUM-SQUARES (ABOUT)	DF	MEAN SQUARE
Treatment (Between)	1	19.0592				
Error (Within)	300	3970.5302	1881.0135	2089.5167	299	6.9884
Treatment + Error (Total)	301	3989.5894	1860.3922	2129.1972	300	
DIFFERENCE FOR TESTING ADJUSTED TREATMENT MEANS				39.6805	1	39.6805

F=5.678    df=1/299

p < .02

TABLE 3.

## SUPPLEMENTARY VARIABLES ON STUDENT RATING FORM

MEASURE (see first appendix for elaboration)	MEANS (s.d.'s below) (n to side)		"t"	p < x (2-tailed)
	CONTROL	EXPERIMENTAL		
Interest in Course	2.27 <sub>(.82)</sub> (165)	2.15 <sub>(.85)</sub> (95)	1.17	.26
% of Reading Done	1.20 <sub>(.52)</sub> (162)	1.06 <sub>(.35)</sub> (95)	2.33	.02
Grade Expected	2.75 <sub>(.87)</sub> (165)	2.85 <sub>(.73)</sub> (95)	.941	.35
% of Classes Attended	1.80 <sub>(.96)</sub> (165)	1.79 <sub>(1.09)</sub> (95)	.078	.9
Workload	3.36 <sub>(.7)</sub> (165)	3.43 <sub>(.75)</sub> (95)	.753	.46
Instruction Geared..	5.0 <sub>(1.76)</sub> (165)	5.02 <sub>(1.70)</sub> (95)	.089	.9
Conditions of Room	3.42 <sub>(.87)</sub> (165)	3.38 <sub>(.77)</sub> (95)	.368	.68
How Much Time & Effort	2.69 <sub>(.88)</sub> (165)	2.04 <sub>(.85)</sub> (95)	5.78	.001



twelve main items on this test could not be considered independent measures, this hypothesis was tested by doing a multivariate analysis on the data using all twelve measures as dependent variables. The hypothesis was soundly rejected. ( $F=.946$ ,  $df=12/166$ ; see Table 4.)

As is obvious, the sample size used for computing this result (179) was substantially smaller than that used for most of the other results in this thesis. This shortcoming was an unavoidable consequence of some incompatibilities between the evaluation form and the computer program on which the multivariate analysis was done. The form scored a "not applicable" answer to an item as an '8', and a "?" answer as a '9'. The program used for the analysis could not differentiate these numbers from true numbers. Therefore, in the analysis of this data, all people who had marked an '8' or '9' on any of the items were removed from the pool. To verify as well as possible that an error had not been made, the multivariate analysis was redone using a new pool of data. In this reanalysis variable 5 (one often found 'NA') was left out of the analysis and only people who had marked an '8' or '9' on one of the other items were removed from the sample. Nonetheless, the reanalysis still did not record any significant difference (see Table 4).

Besides looking at the evaluation in overall terms, two specific items were examined separately. Item 12 (see Appendix 1) might have been expected to show an overall effect, even if the Manova had not. It did not ( $t'=.01$ ,  $p=.5$ ; see Table 5). Item 7, since it related to testing, might also have been expected to show a difference. It also did not ( $t'=1.40$ ,  $p=.08$ , in the wrong direction; see Table 5).

TABLE 4.

SUMMARY OF RESULTS OF MULTIVARIATE TESTS OF SIGNIFICANCE  
 (using Wilks Lambda Criterion)

SAMPLE	F	DFhyp	DFerr	p < x	R
n=179	.946	12	166	.503	.253
results of reanalysis without variable no. 5					
n=222	1.131	11	210	.339	.236

TABLE 5

## BREAKDOWN OF DATA USED FOR MULTIVARIATE TEST ON EVALUATION FORMS

	MEANS - (S.D.'s in parens)		't'	p(t)	p(F)	UNIVARIATE	
	CONTROL n=114	EXPERIMENTAL n=65				F df=1/177	mean square
VAR 01	5.54(1.07)	5.69 (.98)	.91	.19	.18	.849	.912
VAR 02	5.49(1.08)	5.69(.88)	1.27	.10	.10	1.625	1.674
VAR 03	5.91(1.18)	5.85(1.18)	.36	.64	.65	.130	.181
VAR 04	5.59(1.24)	5.52(1.17)	.34	.63	.63	.117	.173
VAR 05	5.45(1.22)	5.49(1.24)	.24	.41	.41	.056	.085
VAR 06	5.85(1.12)	5.94(1.14)	.50	.31	.31	.251	.318
VAR 07	4.59(1.40)	4.28(1.45)	1.40	.92	.92	1.973	3.999
VAR 08	4.43(1.32)	4.54(1.36)	.52	.30	.30	.274	.489
VAR 09	5.09(1.31)	5.17(1.23)	.41	.35	.35	.168	.275
VAR 10	5.79(1.01)	5.82( .95)	.17	.44	.44	.028	.028
VAR 11	5.00(1.40)	5.14(1.32)	.65	.26	.26	.423	.794
VAR 12	5.74(1.26)	5.74(1.25)	.01	.50	.50	.000	.000

VAR 01 to VAR 12 - These are the main questions on the Teacher evaluation form. For their detailed specifications see Appendix 1.

NOTE - All probabilities are figured one-tailed, assuming an advantage to the experimental section.



Two groups of people were identified as tending towards taking more tests. Those scoring high on the internal vs. external control scale (that is, people who feel that they have control over the events in their lives), took more tests (corr.=.23,  $p=.008$ ,  $n=117$ ). However, as the discussion of interaction effects later on in this paper suggests, for this group the opportunity might have been counterproductive.

Women took more tests than men (men: mean=7.378, s.d.=2.28,  $n=37$ ; women: mean=8.286, s.d.=1.92,  $n=80$ ;  $t'=2.2$ ,  $p=.024$ ) and in their case this did lead to higher scores. Women outscored men by 14.8 to 14.4 in the control group, while in the experimental group this margin was raised to 15.9 to 13.7 (see Table 6).

One of the reasons proposed for supposing that increased frequency of test taking, and other features of this innovation, would facilitate better performance by the experimental group, was that it was supposed that the innovation examined here would stimulate a higher level of work on the part of the students. The evidence relating to this point, while of course inconclusive, is somewhat interesting.

It can be said with fair certainty that the people in the experimental group saw themselves as having done more work than people in the control group. Although they rated the "workload" of the course no differently than the control group (see Table 7) they reported having put more "time and effort" into the course (low no.= more work, control=2.69, experimental= 2.04;  $t'=5.78$ ,  $p<.001$ ; see Table 7), they reported a higher percentage of assigned reading done (low no.=more; control=1.20, exp.=1.06;  $t'=2.33$ ,  $p=.02$ ) and they reported having worked more hours

TABLE 6.

## BREAKDOWNS BY SEX

## OPTIONAL FINAL SCORE by GROUP and SEX

SEX	MEANS (s.d.'s in parens)				't'	p < x
	CONTROL	n	EXPERIMENTAL	n		
Men	14.42 (4.22)	65	13.68 (3.67)	37	.882	.2
Women	14.79 (3.5)	121	15.91 (3.1)	80	2.31	.01

## TOTAL SCORE IN COURSE by GROUP and SEX

SEX	MEANS (s.d.'s in parens)				't'	p < x
	CONTROL	n	EXPERIMENTAL	n		
Men	105.94 (17.0)	65	101.19 (20.03)	37	1.26	.11
Women	107.74 (15.81)	121	107.94 (11.82)	80	.101	.5

TABLE 7.

## MEANS OF VARIABLES RELATING TO DISCUSSION OF WORK

VARIABLE	MEANS (s.d.'s in parentheses)				't'	p < x
	CONTROL	n	EXPER.	n		
Workload (Lo is light)	3.36 (.7)	165	3.43 (.75)	95	.753	.46
Time and Effort (Lo is more)	2.69 (.88)	165	2.04 (.85)	95	5.78	.001
% of Assigned Reading Done (Lo is more)	1.20 (.52)	162	1.06 (.35)	95	2.33	.02
Work (Hrs./wk.)	2.58 (1.8)	186	5.08 (2.2)	117	10.73	.001



per week in the course than the control group ( 5.08 vs. 2.58 for experimental vs. control respectively;  $t=10.73$ ,  $p<.001$ ; Table 7).

Unfortunately, it is difficult to connect this extra perceived effort with better performance. In the control group, score on the optional final correlated significantly with "work" reported done (corr.=.18,  $p=.007$ ; see Table 8 for all correlations which follow), whereas in the experimental group the correlation was not as high (corr.=.13) and was consequently not statistically significant. In neither group was the correlation of total score in the course and "work" statistically significant (corr.=.07 for the control group and corr.=.10 for the experimental group;  $p=.19$  and  $p=.13$ ).

#### Aptitude Treatment Interactions

The analyses of the aptitude treatment interactions (ATI's) in this study were done using a computer program called ANALATI (Dowaliby, 1972). This program provides a test of parallelism between the slopes of the regression lines for the correlation between each trait and each criterion measure and gives an exact probability for each test of parallelism. It also does a Johnson-Neyman test on the data, a test which determines a region of non-significance around the cross points for any chosen level of statistical significance. Beyond that region, it can be said with some authority that a subject would benefit by a particular treatment.

Two interaction effects were noted in this experiment. The first concerns test anxiety. It was found that people who were very test

TABLE 8.

## CORRELATIONS OF SOME SELECTED VARIABLES

GROUP	VARIABLES	CORRELATION	DF	$p < x$
Control	Optional final score with Work	.18	184	.007
Exper.	Optional Final Score with Work	.13	115	.09
Control	Total Score with Work	.07	184	.18
Exper.	Total Score with Work	.10	115	.13
Exper.	Total Number of Tests with Work	.11	115	.12
Exper.	Optional Final Score with Total No. of Tests	.16	115	.05
Exper.	Total Score with Total No. of Tests	.22	115	.009
Exper.	Optional Final Score with Total No. of Tests (controlling for Work)	.14	114	.06
Exper.	Total Score with Total No. of Tests (controlling for Work)	.21	114	.012

anxious, as measured by the Wiggins short form of the test anxiety scale, performed better on the optional final if they had been in the experimental condition (low=less than 4.92; see Figure 1). It was also seen that a student with low test anxiety, similarly measured, (more than 8.17; see Figure 2) would be expected to have a higher total course score if he were placed in the control group. These results depend almost entirely on the large effect exerted by test anxiety on the control group students (corr.=.33 and corr.=.38 for test anxiety (ANX) with optional final score (OPF) and total course score (TOSC) respectively; see Table 9 for details of all correlations listed in ATI section). It seemed that in the experimental group, anxiety had little or no effect (corr.=.09 and corr.=.12, ANX with OPF and TOSC respectively;  $p=.16$  and  $p=.10$  respectively).

Inspection of the data revealed that in the control group, optional final score and total course score were both highly related to grade point average (GPA) (corr.=.46 and corr.=.54 respectively) and that grade point average was, in turn, significantly related to anxiety, (corr.=.23,  $p<.001$ ). To guard against a spurious relationship, the correlations of anxiety in the control group were redone with the factor of grade point average partialled out. This recalculation did not markedly affect the initial results (corr.=.26 and corr.=.32 for ANX with OPF and TOSC respectively; both have  $p<.001$ ).

One must be cautious about drawing conclusions about "test anxiety". The other scale used, the Anxiety Differential (AD), a more widely used



TABLE 9.

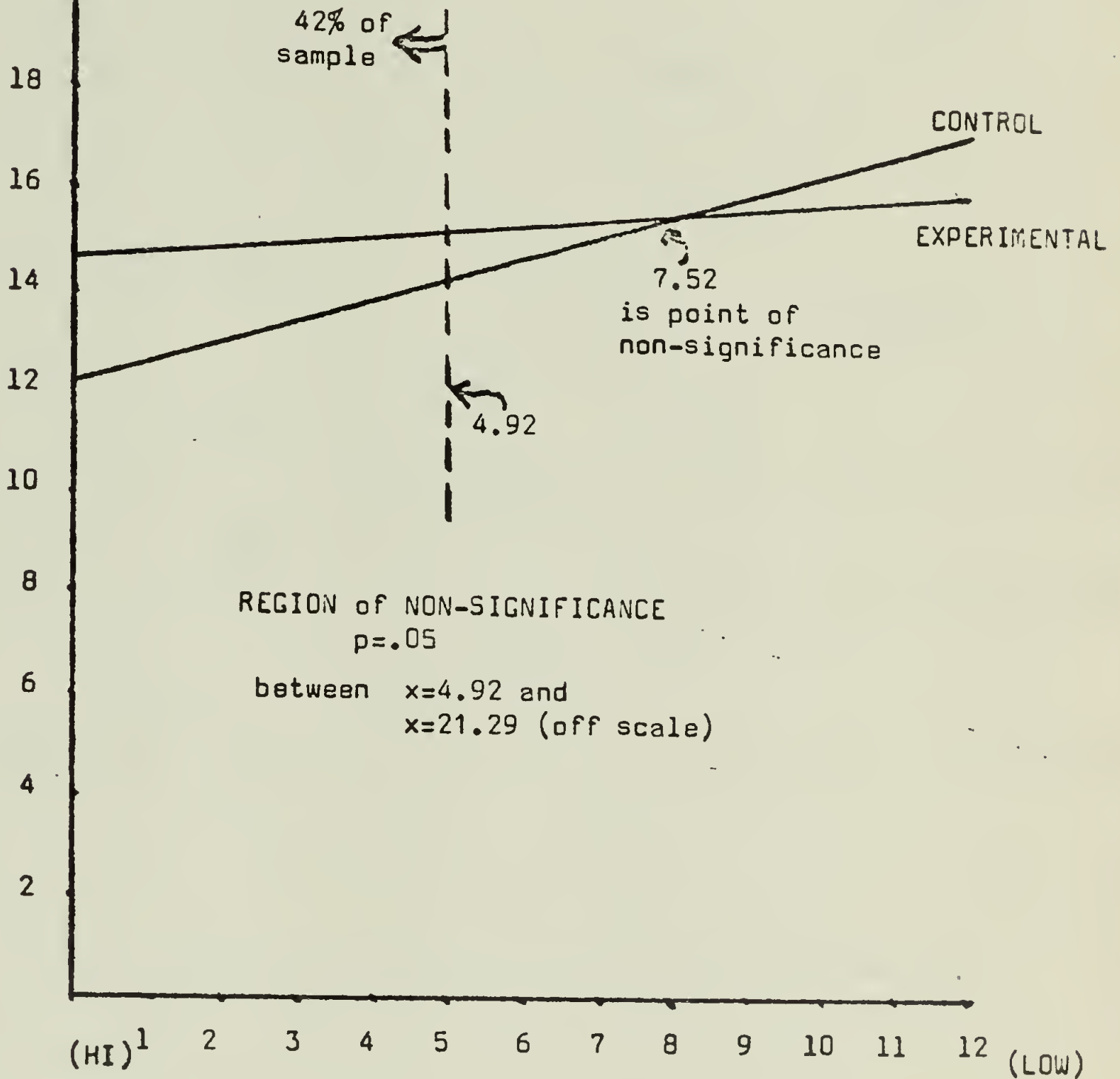
## CORRELATIONS USED IN DISCUSSION OF INTERACTIONS

GROUP	VARIABLES	CORRELATION	DF	p < x
Control	Optional Final Score with Anxiety Short Form	.33	184	.001
Control	Total Score with Anxiety Short Form	.38	184	.001
Exper.	Optional Final Score with Anxiety Short Form	.09	115	.16
Exper.	Total Score with Anxiety Short Form	.12	115	.10
Control	Optional Final Score with Grade Point Average	.46	184	.001
Control	Total Score with Grade Point Average	.54	184	.001
Exper.	Optional Final Score with Grade Point Average	.45	115	.001
Exper.	Total Score with Grade Point Average	.47	115	.001
Control	Anxiety Short Form with Grade Point Average	.23	184	.001
Control	Optional Final Score with Anxiety Differential	-.13	184	.04
Control	Total Score with Anxiety Differential	-.13	184	.04
All	Anxiety Short Form with Anxiety Differential	.30	302	.001
All	Optional Final Score with SAT -Verbal	.30	302	.001
All	Optional Final Score with Grade Point Average	.45	302	.001

FIGURE 1.

SCORE ON  
OPTIONAL  
FINAL

Regression Lines of Experimental vs. Control on  
Optional Final Score by Anxiety



EQUATION OF REGRESSION LINE

CONTROL -  $y = 12.23 + .42x$

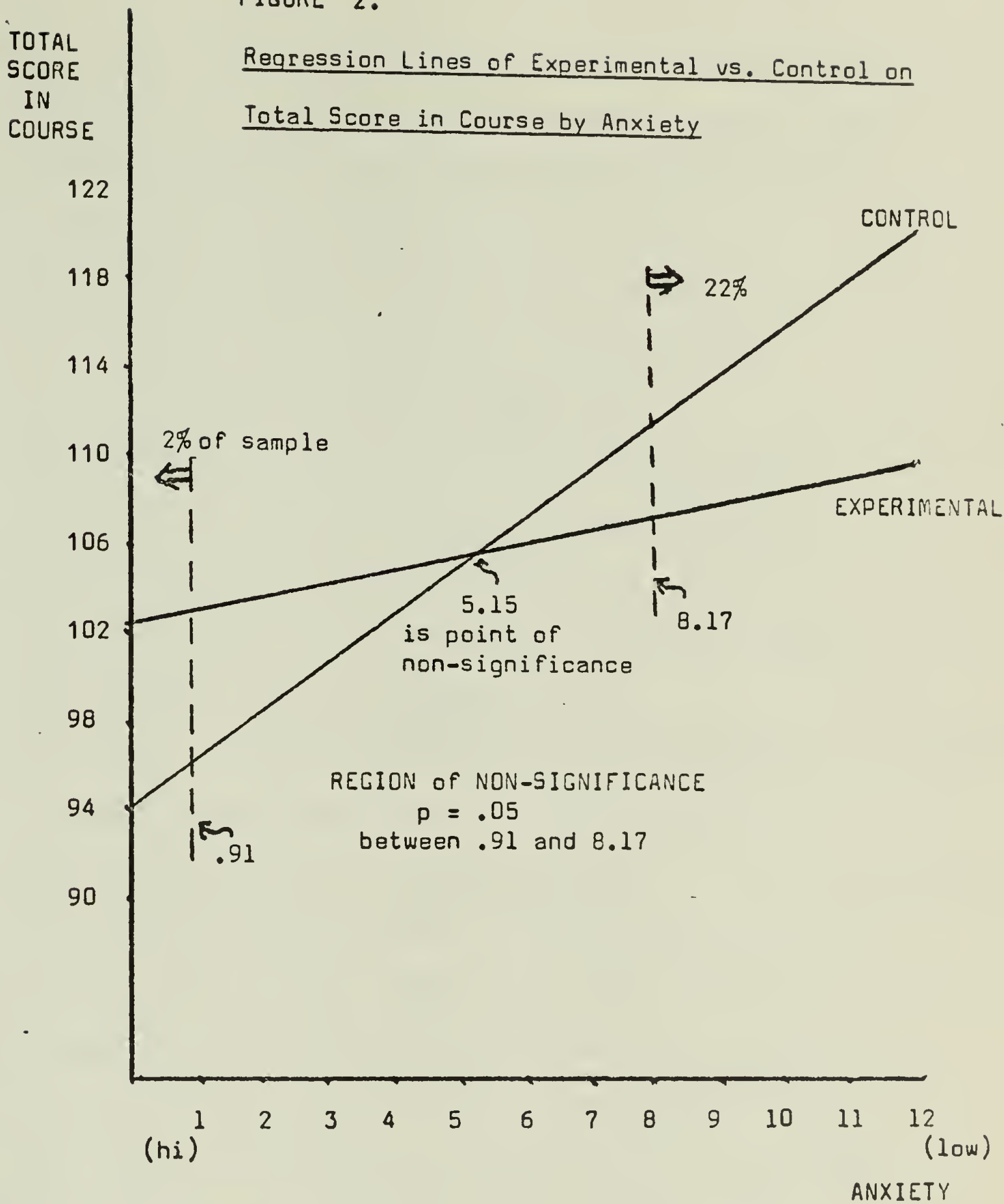
EXPER. -  $y = 14.62 + .11x$

TEST of COMMON SLOPE

$F(1,299) = 5.0812$

$p = .0234$

FIGURE 2.



EQUATION OF REGRESSION LINE

CONTROL -  $y = 94.5 + 2.15x$   
 EXPER. -  $y = 102.5 + .6x$

TEST of COMMON SLOPE

$F(1,299) = 6.3712$   
 $p = .0117$

scale, displayed a much lower relationship with optional final score and total score in the course in the control group than did the Wiggins short form just discussed (corr.=.13,  $p=.04$  for the anxiety differential with both optional final score and total course score) and displayed an equally low relationship as the short form did in the experimental group (corr.=-.07 and corr.=.03 for the anxiety differential with optional final score and total course score respectively.)

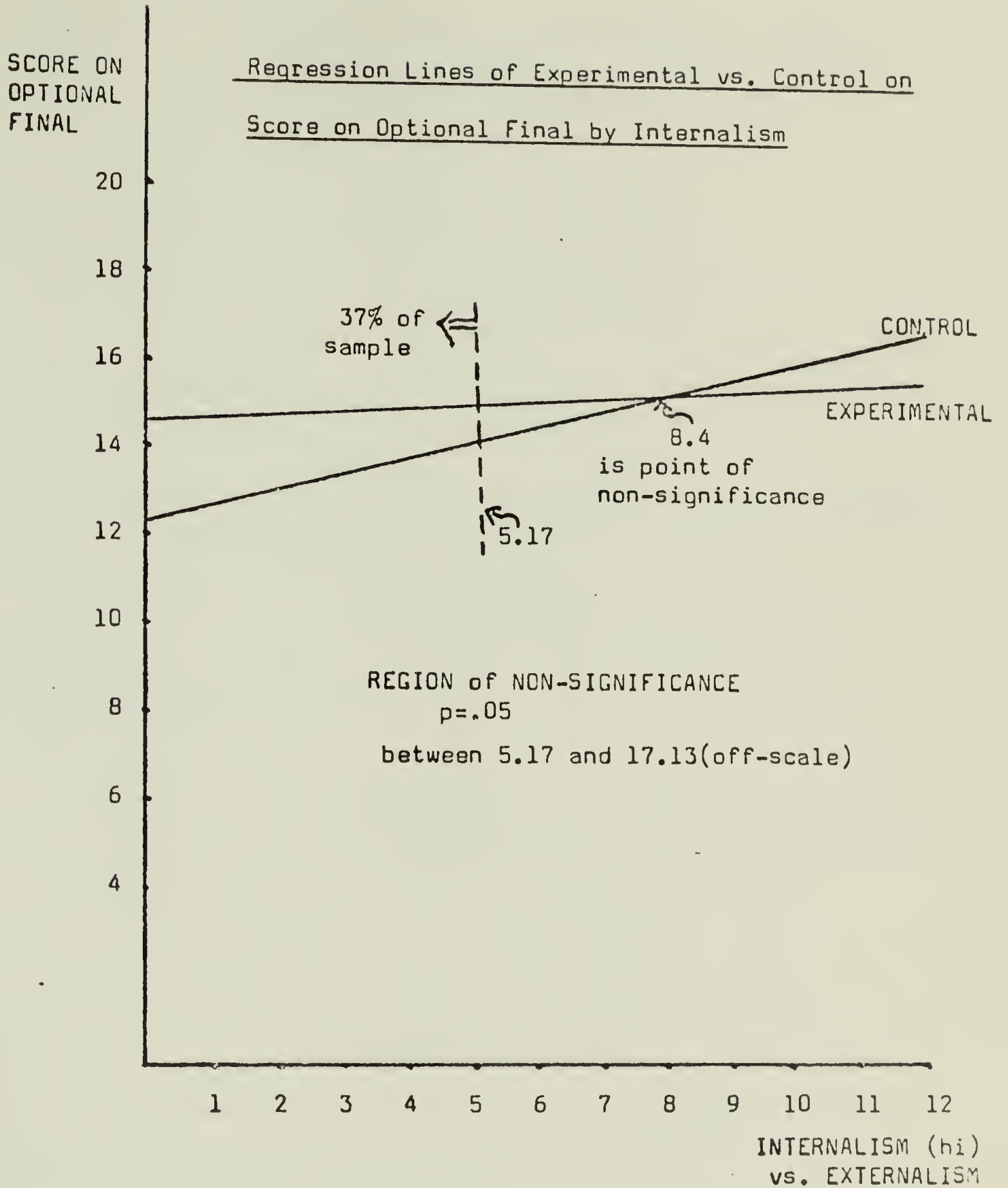
This lower relationship in the control group caused the ANALATI program to accept the null hypothesis that the regression lines of the two groups for both optional final score and total course score were not really different in slope (test of common slope : for OPF  $-F=.3163$ ,  $df=1/301$ ,  $p=.58$ ; for TOSC  $-F=2.1986$ ,  $df=1/301$ ,  $p=.14$ ).

It should be noted that these differences in predictive worth between the two scales occurred despite the fact that the two scales were significantly correlated (corr.=.3,  $p=.001$ ,  $n=303$ ).

The other interaction studied yielded a counterintuitive yet significant result (see Figures 3 and 4). While it was clear from the data that students judged to be internally controlled, took more advantage of the innovation, it seems clear that they did not benefit by it. Looking at optional final score, with reasonable assuredness ( $p=.05$ ) we can make the statement that being in the experimental group benefitted those students more external than internal (less than 5.17). However, looking at total course score, we can say (again  $p=.05$ ) that those students external in control (less than 2.27) benefitted by being in the experimental group, but a larger number of internal students benefitted by being in the control group.



FIGURE 3.



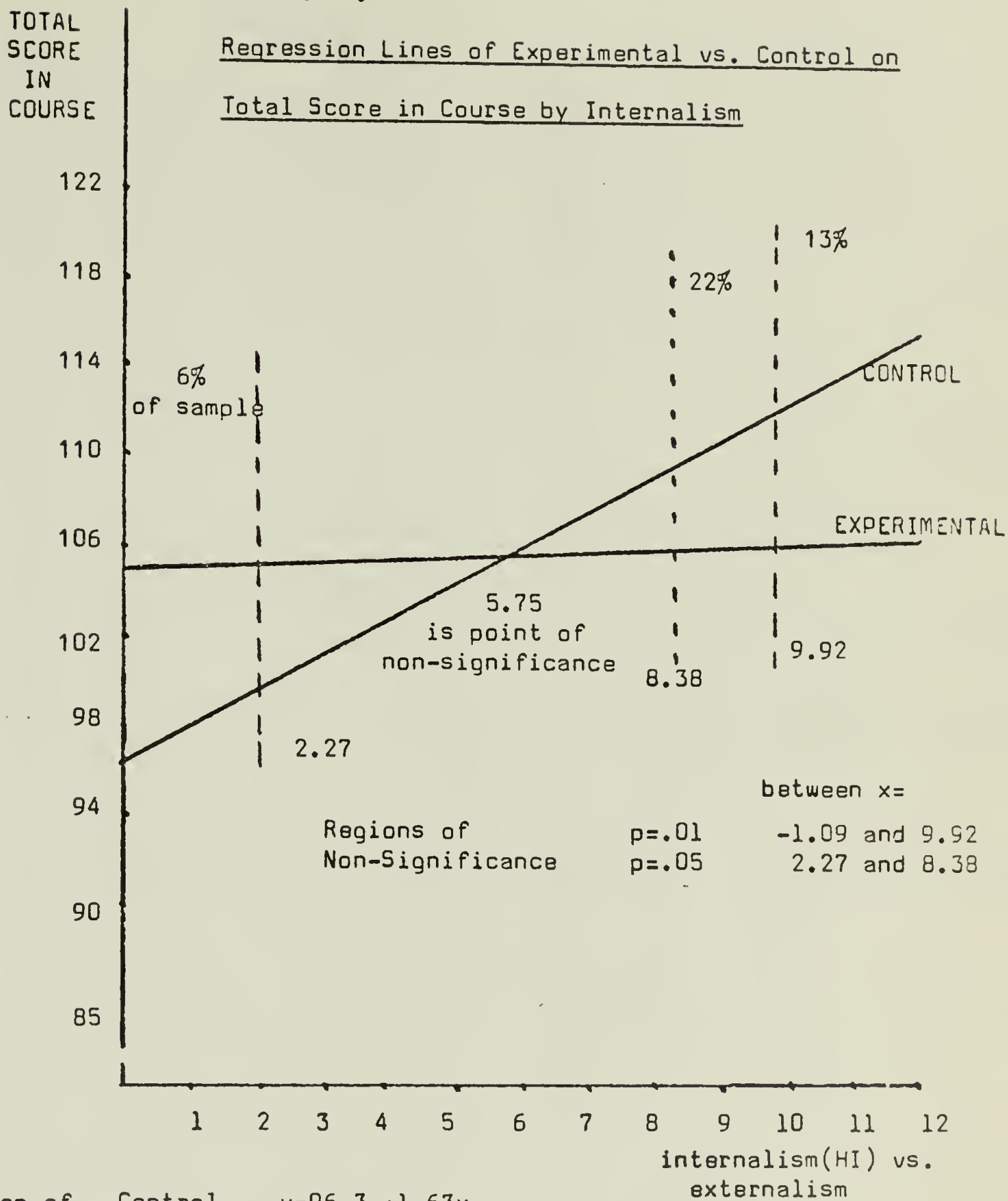
EQUATION OF REGRESSION LINE

CONTROL -  $y = 12.32 + .36x$   
 EXPER. -  $y = 14.55 + .09x$

TEST of COMMON SLOPE

$F(1,301) = 6.34$   
 $p = .0119$

FIGURE 4.



Test of COMMON SLOPE

$F(1,301) = 10.4733$        $p = .0017$

## DISCUSSION

It can be said that the significant difference attained in the results between the means of the two groups on their optional final score proves that this innovation is, on the average advantageous to all concerned. However, a more critical appraisal must be that the small gains in learning and attitude do not recommend the widespread adoption of this procedure.

It should be noted here that casual class reaction to the experimental procedure itself was quite favourable. A surprising number of students commented to both the professor and his assistant that the innovation was a remarkable improvement. This was not reported in the previous section for three reasons. First, it can easily be attributed to a 'Hawthorne' effect. Second, it is clearly not a random or representative sample of opinion. And third, casual reaction being favourable is a trivial result, there being no disadvantage to being in the experimental section, and therefore no rigorous tests of the finding were prepared. However, this casual student reaction was, by itself, favourable enough to suggest to the instructor that he should continue the innovation as the norm in future presentations of the course. This decision must be reevaluated in light of the data.

For the reason expressed above, it was very surprising that there were no differences between the groups on the evaluation instrument. The result may stand in testimony to the fact that students are, on the whole, even handed judges of teaching quality and are not swayed by the little things attached only peripherally to the course content.

Of course, that there were no differences on the evaluation and that the learning gains were so small might be the result of some of the limitations inherent in the execution of the experiment rather than in the innovation itself.

The largest flaw in the experiment was that, by limiting the optional final to material covered in the first half of the course, the measures did not tap what was probably a gradually increasing gain. It is likely that students were more fully adapted to the innovation and its advantages only towards the end of the semester. Unfortunately, no data was taken that might show increasing utilization of the options as the semester proceeded. Were the experiment to be repeated, it would be worthwhile to make an effort to attain a true post-test score, one taken three or four months after the end of the course.

The items used in the tests of the material deserve some comment for they may be partly to blame for the minimal results achieved in the experiment. Most people would say that final exams, and in particular reasonably short ones which are made up of objective items, are inadequate probes of knowledge gained in a course. This author disagrees with that point. It is felt here that the items were quite adequate as probes of individual knowledge, and even failing that, were more than adequate for assessing group gains. However, it is quite possible that the items were viewed as inadequate by the students, causing the tests to be viewed as largely arbitrary and therefore reducing the effects of any manipulation based on them. Whether it is possible to make tests seem fairer without adding a serious objective specification component



to the course is an open question. However, the whole problem of how to probe the observed fairness of exams and how to improve it is a most worthwhile area for future research and experimentation.

The most interesting results of the study were those associated with the aptitude treatment interaction analyses.

The fact that people high in test anxiety did better in the experimental section speaks well of the innovation. However, the significant fact is that test anxiety barely affected results within the experimental section. While most people show lesser or greater degrees of test anxiety, some people are seriously hindered by it in testing situations. These people are usually treated, with mixed success, using psychiatric techniques such as hierarchical desensitization. However, if we recognize that the negative effects of test anxiety are felt mainly in school situations, that these situations seldom correspond to real-life situations, and that, as is here shown, some test situations can be constructed which will not elicit these negative responses, we see that a more profitable way of helping people high in test anxiety might be to expose them to alternative testing situations that would not hamper their performance. Discovering and refining these situations would be a most fruitful area of future research.

As was noted in the results section, the aptitude treatment interaction analysis on internalism versus externalism yielded a counterintuitive result. It was assumed that those students who feel they have control over their own lives would take more tests, score better on each one, and remember more in the end. Although they did, in fact, take more

tests, neither of the other assertions were borne out.

The fact that they remembered less would not have been hard to explain. One could easily say that since the goal was each test, and not long-term retention, a truly goal oriented person would not remember as much material in the long term. However, the fact that those high in internal control also scored significantly worse in the experimental section on their total score in the course has no logical explanation. This author, at a loss for a rationalization, can only suggest that the study, or a variant on the same theme, be done as a replication of this finding to see whether it can stand up to repetition. If it could it would clearly present a serious explanatory challenge to future theorists.

The implications of this study for educational practice and research are both varied and interesting. The first and most important implication is that the possibility exists of constructing test situations that do not handicap people who have high test anxiety. This possibility alone is important enough to justify significant future research.

The second implication of this study is that in future educational research the factor of sex must be examined. The fact that women reacted favourably to this innovation was a serendipitous finding, but the strength of the effect in the absence of any obvious explanation indicates that individual sex differences should more often be considered a legitimate question in educational research.

The third and final implication is the idea that one must be prepared to substantially restructure existing courses in order to affect serious learning gains. It is clear that the manipulation discussed here was

both too small and too peripheral to the essential business of instruction to affect any clear improvement in learning.

In summary, it can be restated that the major goal of the study, to design an educational innovation that would make a significant improvement to present practice, ended in failure. This particular innovation neither aided student grades, nor student learning, nor did it better the students attitude toward either the course as a whole or the testing facet of the course.

However, certain interesting theoretical points were raised. The finding that sex of student can make a real contribution to the success or failure of a technique was serendipitously discovered. The idea that certain testing situations can neutralize the usually negative effect of test anxiety was advanced with some support. And a counterintuitive effect of internal control was brought out and suggested for future study. All in all, a reasonable venture.

FIGURE 5 .

## TESTS AVAILABLE ON EACH TESTING DAY

		TEST ON THE SEGMENT OF MATERIAL ASSOCIATED WITH WEEK .....											
		1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
TESTING DAY NUMBER		A											
1.		A											
2.		B	A										
3.		C	B	A									
4.		D*	C*	B*	A*								
5.			D	C	B	A							
6.				D	C	B	A						
7.					D	C	B	A					
8.						D*	C*	B*	A*				
9.							D	C	B	A			
10.								D	C	B	A		
11.									D	C	B	A	
12.										D*	C*	B*	A*

\* Starred forms were put together to be the mid-term exams on the indicated days for the control group.



APPENDIX 1.

PLEASE DO NOT WRITE HERE

STUDENT ID NUMBER

8 digit form for student ID number

Grading scale: A (90-100), B (80-89), C (70-79), D (60-69), F (50-59), I (40-49), NA (0-39)

Grade of course: A (Very High), B (High), C (Average), D (Low), F (Very Low)

Class standing: Freshman, Sophomore, Junior, Senior

Grade of this course: A, B, C, D, F, NA

The physical conditions... (handwritten notes about course conditions)

Directions: Please complete the identification information... (instructions for the survey)

INSTRUCTIONS

USE OF PENCIL ONLY - DO NOT USE INK OR BALLPOINT PEN

- 1 = Hopelessly Inadequate, 2 = Inadequate, 3 = Fairly Adequate, 4 = Adequate, 5 = Effective, 6 = Very Effective, 7 = Casually Effective, NA = No Answer, 0 = Don't Understand

1. THE INSTRUCTOR... TO OTHER AREAS OF KNOWLEDGE. 2. THE INSTRUCTOR... SUMMARIZED MAJOR POINTS... 3. THE INSTRUCTOR... PARTICIPATION... 4. THE INSTRUCTOR... STUDENTS HAS A GENUINE INTEREST... 5. THE INSTRUCTOR... INTERESTS, EXPERIENCE OR ABILITY TO GIVE... 6. THE INSTRUCTOR ENJOYS TEACHING... 7. THE INSTRUCTOR... MATERIAL. 8. THE INSTRUCTOR... 9. THIS COURSE HAS... 10. THE INSTRUCTOR'S PREPARATION IS APPROPRIATE. 11. THIS COURSE HAS GIVEN ME... 12. THIS COURSE...

13. 14. 15. 16. 17. 18. 19. 20. 21. 22. 23. 24. 25. 26. 27. 28. 29. 30. 31. 32. 33. 34. 35. 36. 37. 38. 39. 40. 41. 42. 43. 44. 45. 46. 47. 48. 49. 50. 51. 52. 53. 54. 55. 56. 57. 58. 59. 60. 61. 62. 63. 64. 65. 66. 67. 68. 69. 70. 71. 72. 73. 74. 75. 76. 77. 78. 79. 80. 81. 82. 83. 84. 85. 86. 87. 88. 89. 90. 91. 92. 93. 94. 95. 96. 97. 98. 99. 100.

## REFERENCES

- Bloom, B.S. Learning for Mastery. Evaluation Comment (newsletter), UCLA, 1968.
- Born, D.G., Gledhill, S.M., & Davis, M.L. Examination performance and number of student withdrawals in lecture-discussion and personalized instruction courses. In Born, D.G. Instructor Manual for Development of a Personalized Instruction Course. University of Utah, 1971.
- Bostow, D.E., Mawhinney, V.T., Laws, D.R., & Blumenfield, G.J. An Analysis of student studying behaviour as a function of two schedules of testing. Paper presented at AERA, Minneapolis, March 1970.  
Also: ERIC 040-456.
- Dubin, R.T., & Taveggia, J. The Teaching Learning Paradox. University of Oregon Press, 1968.
- Fitch, M., Drucker, A., & Norton, J. Frequent testing as a motivating factor in large lecture classes. Journal of Educational Psychology, 1951, volume 42, number 1, 1-20.
- Hertzberg, O.E., Heilmann, J.D., & Leuenberger, H.W. The value of objective tests as teaching devices in educational psychology. Journal of Educational Psychology, 1932, volume 23, 371-380.
- Keller, F.S. "Goodbye Teacher...". Journal of Applied Behavioural Analysis, 1968, volume 1, 79-89.
- Keyes, N. The influence on learning and retention of weekly as opposed to monthly tests. Journal of Educational Psychology, 1934, volume 25, 427-436.
- Kulp, Daniel H. Jr., Weekly tests for graduate students. School and Society, 1933, volume 38, 157-159.
- Kulik, J.A., Kulik, C., and Carmichael, K. The Keller Plan in Science teaching. Science, 1974, 183 (4123), 379-383.

- Mertens, G.C. Student involvement in a relevant social issue - their own education. Mimeographically self-published. 1971, University of Minnesota, St. Cloud, Minn.
- Smeltzer, C.H. An experimental evaluation of certain teaching procedures in educational psychology. Unpublished Doctoral Dissertation, 1931, Ohio State University. In Pressey, S.L. Psychology and the New Education, Harper's, 1933, 363-366.
- Turney, A.H. Effect of frequent short objective tests upon the achievement of college students in educational psychology. School and Society, 1931, volume 33, 760-762.
- Wiggins, J.A., Pope, H., & Bushell, D. Jr. Learning contingencies in the college classroom; a pilot study. June 1968, ERIC 024-314.





