INFORMATION TO USERS

This was produced from a copy of a document sent to us for microfilming. While the most advanced technological means to photograph and reproduce this document have been used, the quality is heavily dependent upon the quality of the material submitted.

The following explanation of techniques is provided to help you understand markings or notations which may appear on this reproduction.

1. The sign or "target" for pages apparently lacking from the document photographed is "Missing Page(s)". If it was possible to obtain the missing page(s) or section, they are spliced into the film along with adjacent pages. This may have necessitated cutting through an image and duplicating adjacent pages to assure you of complete continuity.

2. When an image on the film is obliterated with a round black mark it is an indication that the film inspector noticed either blurred copy because of movement during exposure, or duplicate copy. Unless we meant to delete copyrighted materials that should not have been filmed, you will find a good image of the page in the adjacent frame.

3. When a map, drawing or chart, etc., is part of the material being photographed the photographer has followed a definite method in "sectioning" the material. It is customary to begin filming at the upper left hand corner of a large sheet and to continue from left to right in equal sections with small overlaps. If necessary, sectioning is continued again—beginning below the first row and continuing on until complete.

4. For any illustrations that cannot be reproduced satisfactorily by xerography, photographic prints can be purchased at additional cost and tipped into your xerographic copy. Requests can be made to our Dissertations Customer Services Department.

5. Some pages in any document may have indistinct print. In all cases we have filmed the best available copy.

University
Microfilms
International

300 N. ZEEB ROAD, ANN ARBOR, MI 48106
18 BEDFORD ROW, LONDON WC1R 4EJ, ENGLAND

CURLEY, ROBERT DEAN

THE EFFECT OF VARYING PROPORTIONS OF POSITIVE AND NEGATIVE INSTANCES ON STUDENT MISINTERPRETATION OF STATISTICAL HYPOTHESIS TESTING

The University of Oklahoma PH.D. 1980

THE UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE EFFECT OF VARYING PROPORTIONS OF POSITIVE AND NEGATIVE INSTANCES ON STUDENT MISINTERPRETATION OF STATISTICAL HYPOTHESIS TESTING

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

DOCTOR OF PHILOSOPHY

BY

ROBERT DEAN CURLEY

Norman, Oklahoma

1980

THE EFFECT OF VARYING PROPORTIONS OF POSITIVE AND NEGATIVE INSTANCES ON STUDENT MISINTERPRETATION OF STATISTICAL HYPOTHESIS TESTING

APPROVED BY

Harold V. Hinsche

Larry E. Jackson

J. W. Higgins

Thomas J. Hill

Herbert R. Hengst

DISSERTATION COMMITTEE

ACKNOWLEDGEMENTS

I wish to thank the members of my doctoral committee for their time and efforts on my behalf. Special thanks go to the committee chairman, Dr. Harold V. Huneke, for his advice, patience, and understanding. I feel deeply grateful to Dr. Larry E. Toothaker for the endless sharing of his time, talents, and professional expertise. Dr. Toothaker's encouragement and guidance were a constant source of inspiration to me. It was also a privilege to have worked with Dr. Thomas Wiggins, Dr. Herbert Hengst, and Dr. Thomas Hill.

My thanks go to Terri Higgins for assistance in the preparation of a reading copy of my dissertation via TSO. I extend deep personal thanks to Mrs. Dawn Miller, an extremely dedicated and efficient typist, who worked long hours in retyping the reading copy to bring it to the standards of the Graduate College for a final dissertation copy.

TABLE OF CONTENTS

THE EFFECT OF VARYING PROPORTIONS OF POSITIVE AND NEGATIVE INSTANCES ON STUDENT MISINTERPRETATION OF STATISTICAL HYPOTHESIS TESTING

CHAPTER I

INTRODUCTION

The correct interpretation of statistical terminology and results is an objective often listed for non-calculus courses in elementary statistics (Campbell, 1974; CUPM, 1972; Huck, Cromier, & Bounds, 1974; Johnson, 1975; Joiner & Campbell, 1976; Mendenhall & Ott, 1976). The panel on statistics of the committee on the undergraduate program in mathematics, in its June 1972 report, called: Introductory Statistics Without Calculus, recommends:

> The primary objective of the introductory statistics course should be to introduce students to variability and uncertainty, and to some common concepts of statistics; that is, to methods such as point and interval estimates and hypothesis testing for drawing inferences and making decisions from observed data. (p. 5)

With these objectives in mind, and with the increasing reliance of the sciences on statistical methods as a means of inferring truth, it seems that consumers, as well as producers of research, should have the ability to appropriately interpret research applications of statistics.

As suggested by Joiner and Campbell (1976), it is highly desirable that the student be patiently observant, fully aware, and both sensitive and sympathetic to the inadvertant errors of the researcher, which may place limitations on one's conclusions. Schulte (1979) points out, and others, i.e., Huff (1954), Kirk (1972), Kruskal (1978), Linton, Gallo, & Logan (1975), and O'Brien & Shapiro (1968) have suggested that there is a considerable problem of misinterpretation in statistics among students, and other consumers of statistical information.

In commenting on O'Brien and Shapiro (1968), Kirk (1972) noted that, "There appears to be a natural tendency among students to attribute surplus meaning to the concept of statistical significance. . . the fact that a test statistic is declared significant tells an experimenter nothing regarding the magnitude of the treatment effects or the practical importance or usefulness of the results." (p. 109)

The present author feels that evidence for one type of misinterpretation can be found in the results of Rosenthal & Gaito (1963), and Beauchamp & May (1964). Their research involved university graduate students and professors who displayed more confidence in a rejected null hypothesis for a sample size of 100 than for a sample size of 10.

With respect to the opposite concern of exaggerating the meaning of failure to reject the null hypothesis, Fisher states: "The null hypothesis is never proved or established, but is possibly disproved in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis." (p. 16)

O'Brien & Shapiro (1968) point out that there are basically four types of findings in an experiment:

1. Small differences exist, and they are not statistically significant.
2. Small differences exist, but they are statistically significant.
3. Large differences exist, but they are not statistically significant.
4. Large differences exist, and they are statistically significant. (p. 674)

There would likely be less misinterpretation of statistical hypothesis testing if #1 and #4 above were the only possible findings. The present author's experience is that the existence of #2 and #3 leads some students to feel that hypothesis testing is meaningless, a situation which, in the opinion of the author, should be corrected.

The present research is concerned with misinterpretation of statistical hypothesis testing. For purposes of this study, a student will be considered to have misinterpreted a hypothesis test when one either overgeneralized or undergeneralized the meaning of a statistical test. Overgeneralization will be said to occur when a subject identifies an inappropriate interpretation as appropriate; undergeneralization, as identifying an appropriate interpretation as inappropriate. In particular, the study focuses on misinterpretation of the two independent sample t test when $\overline{X}_1 > \overline{X}_2$ in the following situations:

1. When $P < .05$ in a very high power test.
2. When $P < .05$ in a very low power test.

3. When $P \geq .05$ in a very high power test.

4. When $P \geq .05$ in a very low power test.

One source of insight into misinterpretation may be found in the work of learning theorists studying the role of using varying proportions of positive and negative instances of a concept. One possible use of positive and negative instances lies in the extinguishing of overgeneralization (Klausmeier, Ghatala, & Frayer, 1974; Shumway, 1971, 1974; Tennyson, Wooley, & Merrill, 1972).

Klausmeier, et al. (1974) cite research by Swanson (1972) which demonstrated that students exposed to negative instances tended to:

1. Show more discrimination, i.e., less overgeneralization than students who did not have such exposure, and
2. Have greater ability to recognize new instances, i.e., less undergeneralization.

The above cited research by Swanson (1972) refers to a situation involving initial learning of the environmental concepts of population, habitat, and community, where no definition of those concepts was presented to the sixth graders who served as subjects for the experiment.

In a replication (Swanson, 1972) which differed from the above study only in that a definition of the concepts involved was presented, there were no significant differences between the experimental groups and the control group with respect to overgeneralization, or undergeneralization. In terms of knowledge of definition of the concepts, students who were given a rational set of positive instances only performed significantly better at the .05 level than students given a rational set of both positive and negative instances.

According to Feldman (1972) when the concepts to be studied were changed from the environmental concepts of population, habitat, and community to the geometric concepts of bilateral, rotational, and translational symmetry, with no definition of the concepts given, there was no significant difference among experimental groups in terms of undergeneralization. However, all groups performed significantly better than the control group. In terms of overgeneralization, the group exposed to a rational set of both positive and negative instances, and the group exposed to two positive instances only for each concept performed significantly better than the control group.

Feldman replicated his own study again, this time giving concept definitions to the subjects. The condition involving a rational set of both positive and negative instances, and the condition involving two positive instances only, were significantly better than the control group in terms of discrimination, i.e., resistance to overgeneralization. The group having a rational set of positive instances only did not differ significantly from the control group in terms of overgeneralization. In terms of generalization, i.e., resistance to undergeneralization, each of the experimental groups (a rational set of both positive and negative instances, a rational set of positive instances only, and two positive instances only) performed better, on the average, than did the control group. Each difference was significant ($P < .05$), but there were no significant differences among the experimental groups.

In terms of knowledge of definition of the concepts, the same pattern was followed. When the criterion was knowledge of interrelationships, the group having two positive instances only did best, and was

significantly better than the control group ($P < .05$). The next best was the group having a rational set of both positive and negative instances. The third best was the group having a rational set of positive instances only; and poorest of all, the control group.

Based on these results, and other research, e.g., Frayer (1970), Klausmeier, et al. (1974), it is concluded that when no concept definition is presented, a rational set of both positive and negative instances is preferable, but that when a concept definition is provided, the type of instance given, and their number, is less relevant.

Bruner (1956), and Hoveland & Weiss (1953) show preference for positive over negative instances. Clark (1971), in a review of literature on positive over negative instances, summarizes the vast majority of the research as being much more favorable to the use of positive instances only. However, Markle & Tiemann (1969), and Tennyson, Wooley, & Merrill (1972), for example, have done research demonstrating the crucial role that a mixture of positive and negative instances can play in the learning sequence.

Markle & Tiemann (1969) suggested that using the proper number and kind of positive instances discourages undergeneralization, and using the proper number and kind of negative instances discourages overgeneralization.

Tennyson, Wooley, & Merrill (1972) used pairs of positive and negative instances which were matched on dissimilarity in irrelevant attributes, similarity in relevant attributes, and empirically obtained probability of correct identification. On the average, the results indicated a decreased tendency toward overgeneralization.

Separate research by Frayer (1970), Swanson (1972), and Feldman (1972) each illustrate that subjects who have had previous exposure to definitions of the concepts in question benefit when presentations of instances were accompanied by questions which were designed to illustrate the relevant attributes of the instances.

Shumway (1971, 1974) found that, on the average, students exposed to approximately equal numbers of positive and negative instances showed less tendency to overgeneralize than students given positive instances only. In his first study (1971) the subject matter involved concepts from eighth grade mathematics. In his second study (1974) he was concerned with negative instances in mathematical concept acquisition, and their transfer effects between the concepts of commutativity and associativity. It is not clear to the present author that Shumway's results would generalize from the situations which he presented, to the more complex questions involved in avoiding misinterpretation in reading the results of statistical tests of hypothesis.

In a review of research, Bourne and Dominowski (1972) concluded that while positive instances seem more effective for simple conjunctive concepts, use of both positive and negative instances appears to be more useful for more complex concepts.

Bourne and Guy (1968) have suggested that if the class of exemplars, or positive instances, are more homogeneous than the class of nonexemplars, then positive instances may result in faster concept acquisition. If the reverse is true, then negative instances are suggested.

Smoke (1933) is cited (Shumway, 1971; Hoveland & Weiss, 1953) as antagonistic to negative instances. Smoke's study is concerned with the

relative merits of two methods of concept learning when serial presentation of the learning material is used. However, Smoke concluded, in italics: "Although negative instances may not make for rapidity in learning they tend to make for accuracy, expecially in the case of difficult concepts." Smoke also points out that students tend to come to wrong conclusions and make "snap judgements" less frequently when taught with a mixture of both positive and negative instances than when positive instances only are employed. He further states: "It appears that in so far as negative instances assist concept learning, they do so largely because of the way in which they prevent the learner from coming to one or more erroneous conclusions while he is still in the midst of the learning process." (p. 588)

It seems fair to summarize past learning theory research in the area of positive and negative instances as having mixed conclusions. It may be that the situation involving positive and negative instances is concept specific.

Mathematicians, in teaching advanced classes of their discipline, have often found negative instances to be a useful teaching aid, as attested by the wide usage of Gelbaum & Olmsted's Counterexamples in Analysis (1964), and Steen & Seebach's Counterexamples in Topology (1968).

Negative instances of appropriate statistical interpretation are used in some statistics texts (Campbell, 1974; Johnson, 1975; Mendenhall & Ott, 1976; 100 modules of Lefkon, Fletcher & Derderian, 1975). Campbell devotes his entire book to the subject, as does Huff (1967). An entire section (Good, 1978) of the International Encyclopedia of Statistics (Kruskal & Tanur, 1978) is also devoted to fallacies in statistical

thinking. However, the author was unable to locate any experimental research dealing with the effectiveness of negative instances in extinguishing student tendency to misinterpret, overgeneralize, or overinterpret statistical results.

The present research attempts to make a small advance into this void. Thus, the author deals with the question, "What is the effect of varying proportions of positive and negative instances, i.e., appropriate and inappropriate interpretations, on student misinterpretation of the meaning of statistical hypothesis testing?"

While one may gain some insight from learning theory research, such as cited above, the present research differs rather markedly in several aspects:

1. The concepts in the present study are more complex.
2. The criteria of the present study is not speed of acquisition, but rather, lack of tendency to misinterpret.

As Smoke (1933) has suggested, negative instances, "tend to make for accuracy, especially in the case of difficult concepts."

It is conjectured that, with increased proportions of negative instances, up to two thirds negative, sensitivity to misinterpretation will increase.

CHAPTER II

SUBJECTS

The 55 students who participated in the study were volunteers from two sections of an elementary statistics course for non-majors at the University of Oklahoma. One section met at 10:30 a.m., and the other at 11:30 a.m., on Monday, Wednesday, and Friday during the spring 1977 semester. Three of the subjects were freshmen; sixteen each were sophomores, juniors, and seniors, respectively; two were graduate students, and two were unclassified. This group of 55 volunteers was considered as one class for the purpose of recording results of the experiment.

Actually, there were three more volunteers than the 55 subjects who participated in the study. However, the score of one subject was deleted from the analysis because of late arrival. Also, scores of two volunteers who completed the study were randomly selected for deletion, in order to bring about a more nearly equal number of subjects in each cell. Equality of cell size is helpful in terms of robustness of violations of assumptions to the model (Glass, Peckham, & Sanders, 1972).

The conducting of experiments had been discussed in class, and students were advised that perhaps a good way to understand, and get a

feel for an experiment, would be to participate in one. They were asked to volunteer for an experiment which would deal with the learning of selected concepts in statistics. The one non-volunteer did a computational exercise instead. Students received the same number of points toward their final grade for either their participation in the experiment, or the alternate assignment.

In addition to discussing the conducting of experiments, the class studied, all in the context of one sample tests, the process of statistical hypothesis testing with emphasis on the meaning of power; types I and II error; alpha; the effect of sample size and variance on power when alpha is held constant; the appropriate interpretation of the one sample t test in relation to power; and other related topics.

Perhaps one index of the subjects' mastery of the concept of power in one sample tests can be found in the high proportion of correct answers to some questions from a quiz given the week prior to the experiment. Those questions appear in the appendices. Out of a total of 55 subjects in the study, the number of times various questions were missed are as stated:

Question	Times Missed
#1	3
#2	0
#3	1
#4	1

Some subjects may have generalized from the one sample concept of power to the two sample power concept without the aid of experimental treatments.

The subjects had also been introduced to both large and small, matched and independent two sample tests.

CHAPTER III

DESIGN

Two variables, one a treatment variable and the other a blocking variable, were arranged in a 2X4 factorial design. The total group of subjects was divided into halves on the basis of whether the student was above or below the combined class median in terms of performance on the first five quizzes and the midterm. Within each half of the combined group students were randomly assigned to one of four treatments in such a way that there were seven observations within each treatment for each half of the class. Treatment I subjects received one third positive instances and two thirds negative instances for each concept. Treatment II subjects received two thirds positive, and one third negative instances for each concept. Treatment III subjects received all positive and no negative instances for each concept. Treatment IV was used as a control group. The subjects in this group took the criterion test immediately on entering the room while the other subjects were using the experimental treatment materials. On completion of the criterion test, those subjects in the control group did one of the treatment exercises, while the subjects in the treatment groups did the criterion test.

Since there were 29 students in each half of the total sample, the random assignment of seven students to each of the four groups left one extra student in each half. These extra students were assigned to a group, but their scores were not included in the analysis. Also, one student in the "above median" control group was unavoidable detained, and therefore was late in arriving. This score was deleted from the analysis. Thus, there were only six observations for the "above median" control cell, but seven observations for each of the other cells.

In all, there were four levels of a treatment variable, and two levels of the cumulative achievement blocking variable.

The criterion test was a teacher-made fifteen item quiz which was designed to measure student sensitivity to misinterpretation, and interpretation of results of the two independent sample t test in situations where $P < .05$ in both very high power and very low power tests; and situations where $P \geq .05$ in very high power and very low power tests.

CHAPTER IV

MATERIALS AND PROCEDURES

The experimental treatments consisted of a four page packet, followed by a two page criterion test. All materials used in the study were teacher made.

Each page of the experimental packet contained 12 statements, each of which required circling an "A" for agreement, or "D" for disagreement, to the right of each statement. The correct answer was typed to the left of each statement, and each correct answer was covered with a small, removable label, or tab.

In order to assure protection of these correct answers from impairment by the tabs to be placed over them, transparent tape was applied. Then, a small white tab was placed over each correct answer, and modified so that subjects could easily remove the tab and adequately receive immediate reinforcement. A band of blue ink through the middle area of the white tab concealed the letter below it. The tabs were then about one third white area which should hold fast to the paper; one third blue, middle area, covering the answer; and one third white, finger-hold area which should hold firmly, except for its edge.

A label similar to the tab, but larger, was used on the front page of each packet as a name plate for each subject in the experiment. The packets were then arranged in separate folders, according to the group to which each subject had been randomly assigned.

At the scheduled time for the experiment, packets were passed out, and subjects were requested to read directions on the front page, but not to open packets until so directed.

The front page of each experimental treatment contained the following brief directions to the subject on how to proceed throughout the exercises:

"Read each of the statements on the following pages. Circle "A" if you agree, and "D" if you disagree. Each time you have circled your choice, then raise the tab at the left to see the correct answer. Many thanks for your cooperation."

On the upper half of page one of the exercises, the subject reads:

"Assume all assumptions are met for the two independent sample t." Then follows the passage:

"Suppose that we have a two tailed two independent sample t test which is highly insensitive to differences in population means (μ's), i.e., a low power test. If we accept the null hypothesis at the .05 level, with $\bar{X}_1 > \bar{X}_2$, this suggests that. . . ."

In the six different interpretations which follow, the subject is asked to indicate agreement or disagreement by circling the "A" or "D" at the right of the statement. After indicating a choice on a statement, the subject is to lift the seal to the left of that statement in order to find the correct response. The subject should then proceed to the next statement.

On the lower half of page one is a description which suggests the same situation as above, but with a slightly different approach. Instead of specifically telling the subject that this situation has a low power test which would imply insensitivity to differences in population means, the subject is informed that there are four observations in each treatment, and that the variance is extremely large. The same six proposed interpretations as in the above item follow in reverse order, and the subject proceeds as directed.

The second page of the exercises follows the same format, but this time a high power test is used with the null hypothesis accepted. The third page has a high power test with the null hypothesis rejected, and the last page has a low power test with the null hypothesis rejected.

Copies of the experimental treatment appear in the appendices.

For Treatment I subjects, two thirds of the interpretations for each situation on each page were negative instances of correct interpretations and one third were positive instances. For Treatment II subjects, one third of the interpretations were negative instances, and two thirds were positive. For Treatment III subjects, all of the interpretations were positive, that is, correct, or appropriate. Treatment IV, the control group, took the criterion test while the other subjects were doing the exercises. The control group had not seen any form of the exercises. When all subjects had completed their task, those who had taken the experimental treatments were given the criterion test.

The criterion, which appears in the appendices, was a fifteen item multiple rank choice quiz, taken immediately after the experimental treatments were completed. In the present study, the multiple rank

choice test required the subject to place a "3" beside the most appropriate response, a "2" beside the next best choice, down to "0" for the worst possible choice. The number of points a subject earned is the number placed beside the correct response. The total score for the test is the sum of points earned for the fifteen questions. The maximum number of points is 45. Multiple rank choice format was used in preference to the usual multiple choice due to increased reliability and simplicity (Poizner, 1974).

The entire experiment, both treatments and criterion test, was conducted within the fifty minute period during which the student normally attends class. Subjects were allowed as much time as desired on both the experimental treatments, and on the criterion test. No subject spent longer than thirty minutes on the experimental treatments, or longer than fifteen minutes on the criterion.

Volunteers from previous mathematics and psychology statistics classes previewed the experiment in order to help the author locate any difficulty or problem which might exist in the exercises, in handling the materials, or in the criterion. They also aided in ascertaining the approximate amount of time needed for each phase of the experiment.

CHAPTER V

THE DEPENDENT VARIABLE

The dependent variable was a criterion test designed to measure many types of misinterpretation of the results of a two independent sample t test. Each mistake the subject makes on this type of test involves a combination of both overgeneralization and undergeneralization. If the subject errs by placing the "3" in the space preceding an incorrect alternative, i.e., misinterpretation, the subject has overgeneralized by identifying a false alternative as being the best answer or interpretation, but has undergeneralized by failing to recognize the correct alternative. The number placed beside the correct answer attempts to estimate the degree of confidence the subject felt in the "true", best alternative, and is, in this sense, related to a measure of resistance to undergeneralization. However, since ties are not permitted (a number may be used only one time for each item), the number placed beside the true best answer, when subtracted from three, tells us the maximum number of times the subject could have overgeneralized (one for each alternative considered).

It could be said, then, that the above mentioned measures of

overgeneralization and undergeneralization are confounded. A variable, defined in terms of two confounded measures, is influenced by either of the two measures acting alone, or by some combination of the two. Misinterpretation is influenced, the author believes, by overgeneralization, undergeneralization, or by some combination of the two acting together.

Thus, the author feels that while it was relevant to discuss some of the literature dealing with use of positive and negative instances, in terms of their influence on overgeneralization, undergeneralization, and knowledge of definition, the reader is reminded that the dependent variable in this study involves misinterpretation of results of the two independent sample t test.

CHAPTER VI

RESULTS

The results of this experiment were in the predicted direction. The ANOVA unweighted means analysis (Yates, 1934; Kirk, 1969; Glass & Stanley, 1970) test for interaction between treatments and achievement did not approach significance ($P = .8329$; $F = .2915$ with 3 and 47 d.f.). This result indicates that, on the average, the amount of difference among treatment means was not greatly different for the "above median" achievers than for the "below median" achievers. In other words, the pattern of observed differences in means was not greatly situation specific, depending upon whether students were above or below the median in performance.

In terms of analysis of differences among treatments, two separate omnibus F tests were conducted. The design originally intended for use in this study involved a 2X4 factorial unweighted means analysis of variance (Glass & Stanley, 1970; Kirk, 1969; Yates, 1934). After this analysis had been completed, Speed & Monlezun (1979) suggested that the unweighted means analysis yields distorted P values, unless the design is a 2^k factorial. The decision was made to check the original

omnibus F test for treatments by collapsing over achievement levels, and carrying out a one way ANOVA on treatments.

The resulting P value for this one way ANOVA was .0112 ($F = 4.095$ with 3 and 51 d.f.) while the original P value for treatment differences in the unweighted means analysis was .0026 ($F = 5.6165$ with 3 and 47 d.f.). The fact that the one way ANOVA yielded a higher P value may result from the decrease in power which would naturally occur by not blocking over performance levels. The basic results still indicate an omnibus F test significant at the .05 level for treatments. Thus, there does appear to be at least one significant contrast among treatment groups.

In terms of power for the sample size used, the maximum probability of detecting a true maximum difference of two standard deviations among treatment means is .99; the minimum probability of detecting a difference of one half standard deviation difference among population treatment means is less than .3. About 19% of the total sample variance can be attributed to treatment differences.

Next, all pair-wise comparisons among treatment means were tested at the .05 experimentwise level using the Aspin-Welch modification of the Tukey Test, as discussed by Games & Howell (1976). This method was used so that the probability of one or more false rejections of the null is held at .05, or approximately at .05, due to possible violation of assumptions (Games & Howell, 1976). Of the six possible comparisons only one was significant by this method. The significant comparison was between the control group and the group receiving two thirds negative, one third positive instances. The P value for this comparison was slightly less than .01. However, all of the six differences among sample means

were in the direction predicted. That is, the group receiving two thirds negative instances did best ($\overline{X}$ = 39.5); the group receiving one third negative instances did next best ($\overline{X}$ = 36.57). Then, the group receiving all positive instances ($\overline{X}$ = 35.64); and finally, the control group ($\overline{X}$ = 31.62).

Further discussion concerning the test used in the analysis which yielded the above results are included in the appendix.

CHAPTER VII

DISCUSSION

The author feels that in spite of limitations, the P values generated in this study are reasonably close to what would have been obtained with a sample of the same size, and randomly selected from undergraduate statistics classes for non-majors in large state supported universities, in the latter half of the 1970's.

Furthermore, the author feels that the study gives limited support to the hypothesis that increased proportions of positive and negative instances, up to two thirds negative and one third positive, yield increased resistance to misinterpretation of statistical hypothesis testing with the two independent sample t test. Also, the study suggests evidence that the above pattern of differences does not depend on whether the subjects were above or below the median performance on previous quizzes and the midterm.

The word "limited" is used for several reasons. First, although the trend was in the predicted direction in terms of sample location measures, only the comparison between two thirds negative and the control group was significant at the .05 level, using the Aspin-Welch modification

of the Tukey Test.

Second, there was only one dependent variable. Other variables, such as retention measures, attitudinal measures, and many others, would also be of interest in future studies. Use of additional variables in the present study would have required a time frame in excess of one 50-minute class period. The author felt that the use of only one class period was necessary in order to insure that the observations be completely independent. Violation of the requirement of independent observations is serious in that it leads to extremely distorted P values, according to research conducted by Box (1954), and reported by Scheffé (1959), Glass et al. (1972), and Elashoff & Elashoff (1978).

Third, subjects did not have experience in computing power for various alternatives and sample sizes, using the two independent sample t test.

Fourth, some subjects may have been demotivated because they had been informed that their performance in the experiment would not affect their grades. The author felt that, for ethical considerations, a subject should not be penalized, either emotionally or academically, because of the treatment to which that student had been randomly assigned.

In terms of reliability, no index was computed, although research by Poizner (1974) shows that this particular grading system, namely the multiple rank choice, maximizes the reliability for multiple choice style tests. The only measure of validity was face validity, as determined by the author, and by Professor Toothaker. Perhaps there are other measures of misinterpretation which might not have lower reliability, but higher validity.

The author does not feel that his results would be significantly affected by violation of underlying assumptions on the basis of voluminous research conducted in this area (Box, 1953; Boneau, 1960; Donaldson, 1968; Elashoff & Elashoff, 1978; Feir & Toothaker, 1974; Games & Lucas, 1966; Glass, et al., 1972; Norton, 1949; Sheffé, 1959; Toothaker, 1972; and many others). Further discussion concerning analysis of data is contained in the appendices.

The author feels that the present study does have implications for further research in the area of the differential effect of positive and negative instances on misinterpretation of statistics, as well as other areas of both mathematics and science. However, the author does not feel that the results of this study have any immediate implications for use in the statistics classroom. The hesitancy felt by the author stems both from the above stated limitations in the present study as well as the lack of related literature which the author was able to locate. Classroom implementation should occur only as a result of sound philosophical examination, as well as an extended series of empirical studies, all, or most of which, tend to support the same, or similar, conclusions. Hopefully this on-going process of examination, and reexamination will continue in statistics education. For the survival of a technical, democratic state which bases important, and, at times, even life saving decisions on the results of statistical tests, a sound education in data analysis and interpretation is vital. Recent events make obvious the increasing reliance of the sciences on statistical methods as a means of infering truth. This dependence of society on statistics implies that both producers and consumers of research should have considerable ability

to interpret research applications of statistics.

Joiner and Campbell (1976) in Modular Instruction in Statistics: Report of the American Statistical Association, state:

> There is little doubt that statistics has become one of the key tools of science, history, and indeed, society as a whole. Today an understanding of statistical concepts is important to every citizen and crucial to anyone doing quantitative research in the social, biological or physical sciences. Yet, statistics remains a widely feared and little understood subject. (pp. 87-88)

BIBLIOGRAPHY

AMSTAT News, American Statistical Association, Washington, D. C., March 1975, Nov. 1973.

Aspin, A. A., "Tables for Use in Comparisons Whose Accuracy Involves Two Variances Separately Estimated," Biometrika, 1949, vol. 36, pp. 290-293.

Aspin, A. A., "An Examination and Further Development of a Formula Occuring in the Problem of Comparing Two Mean Values," Biometrika, 1948, vol. 35, pp. 88-96.

Beauchamp, K. & May, R. B., "Replication Report: Interpretation of Levels of Significance by Psychological Researchers," Psychological Reports, 1964, vol. 14, p. 272.

Binder, A., "Further Considerations on Testing the Null Hypothesis and the Strategy and Tactics of Investigating Theoretical Models," Psychological Review, American Psychological Association, 1963, vol. 70, no. 1, pp. 107-115.

Boneau, C. A., "The Effects of Violations of Assumptions Underlying the t-Test." (Ed.: Lieberman, B., 1971, pp. 357-370.) Original Source: Psychological Bulletin, 1960, vol. 57, no. 1, pp. 49-64.

Boneau, C. A., "Comparison of the Power of μ and t-Tests." Psychological Review, 1962, vol. 69, no. 3, pp. 246-256. (Ed: Lieberman, B., 1971, pp. 175-185.)

Borden, R. C., Individualized Statistical Problems Generating Routine. University of Oklahoma, Dept. of Psychology, 1976.

Bourne, L. E. & Dominowski, R. L., "Thinking," Annual Review of Psychology, 1972, pp. 105-130.

Bourne, L. E. & Guy, D. E., "Learning Conceptual Rules: The Role of Positive and Negative Instances," Journal of Experimental Psychology, 1968, vol. 77, no. 3, pp. 488-494.

Bourne, L. E. & Guy, D. E., "Learning Conceptual Rules: I. Some Inter-rule Transfer Effects," Journal of Experimental Psychology, 1963, vol. 76, pp. 423-429.

Bourne, L. E. Jr., Ekstrand, B. R., & Montgomery, B., "Concept Learning as a Function of the Conceptual Rule, and the Availability of Positive and Negative Instances," Journal of Experimental Psychology, 1969, vol. 82, no. 3, pp. 538-544.

Box, G. E. P., "Non-Normality and Tests on Variances," Biometrika, 1953, vol. 40, pp. 318-335.

Box, G. E. P., "Some Theorems on Quadratic Form Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance on the One-Way Classification," Annals of Mathematical Statistics, 1954, vol. 25, pp. 290-302.

Box G. E. P., "Some Theorems on Quadratic Form Applied in the Study of Analysis of Variance Problems, II. Effect of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification," Annals of Mathematical Statistics, 1954, vol. 25, pp. 484-498.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. A Study in Thinking. New York: Wiley, 1956.

Campbell, S. Flaws and Fallacies in Statistical Thinking. Englewood Cliffs, N. J.: Prentice-Hall, 1974.

Clark, C. D., "Teaching Concepts in the Classroom: A Set of Teaching Prescriptions Derived from Experimental Research," Journal of Educational Psychology, 1971, vol. 62, no. 3, pp. 253-271.

CUPM, (Committee on the Undergraduate Program in Mathematics, Panel on Statistics). Introductory Statistics Without Calculus. American Mathematical Association, June, 1972.

David, F. N. & Johnson, N. L., "The Effect of Non-Normality on the Power Function of the F-Test in the Analysis of Variance." Biometrika, 1951, vol. 38, pp. 43-57.

Donaldson, T. S., "Robustness of the F-Test to Errors of Both Kinds, and the Correlation Between the Numerator and Denominator of the F-Ratio, "Journal of the American Statistical Association, 1968, vol. 63, pp. 660-676.

Elashoff, R. & Elashoff, J., "Effect of Violation of Assumption Errors," International Encyclopedia of Statistics, 1979, vol. 1, pp. 229-250.

Feir-Walsh, B. J. & Toothaker, L. E., "An Empirical Comparison of the ANOVA F-Test, Normal Scores Test and Kruskal-Wallis Test Under Violation of Assumptions," Educational and Psychological Measurement, 1974, vol. 34, pp. 789-799.

Feldman, K. V. The Effects of Number of Positive and Negative Instances, Concept Definition, and Emphasis of Relevant Attributes on the Attainment of Mathematical Concepts. Madison: Wisconsin Research and Development Center for Cognitive Learning. Tech Report, no. 243, 1972.

Fisher, R. A. The Design of Experiments. Edinburgh: Oliver & Boyd, 1949, (5th ed.), p. 16.

Frayer, D. A. The Effects of Number of Instances, and Emphasis of Relative Attribute Values on Mastery of Geometric Concepts by 4th and 6th Grade Children. Madison: The Wisconsin Research and Development Center for Cognitive Learning, Technical Report, no. 116, 1970.

Gagne, R. M. The Condition of Learning. New York: Holt, Rinehart, and Winston, 1965.

Gelbaum, B. R. & Olmstad, J. M. H. Counterexamples in Analysis. San Francisco: Holden-Day, Inc., 1964.

Games, P. A., "Multiple Comparisons on Means," Journal in Research Education, vol. 8, no. 3, pp. 531-565.

Games, P. A. & Howell, J. F., "Pairwise Multiple Comparison Procedures with Unequal N's and/or Variances: A Monte Carlo Study," Journal of Educational Statistics, 1976, vol. 1, no. 2, pp. 113-125.

Glass, G. V., Peckham, P. D., & Sanders, J. R., "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance," Review of Educational Research, vol. 42, no. 3, pp. 237-288.

Glass, G. V. & Stanley, J. C. Statistical Methods for Education and Psychology. N. J.: Prentice-Hall, 1970.

Good, Irving J., "Fallacies in Statistics," International Encyclopedia of Statistics, N. Y.: Free Press, 1978, vol. 1, pp. 337-349, (edited by Kruskal, W. H. & Tanur, J. M.).

Horsnell, G., "The effect of Unequal Group Variances on the F-Test for the Homogeneity of Group Means," Biometrika, 1953, vol. 40, pp. 128-136.

Hoveland, C. I. & Weiss, W., "Transmission of Information Concerning Concepts Through Positive and Negative Instances," Journal of Experimental Psychology, 1953, vol. 45, no. 3, pp. 175-182.

Huck, S. W., Cromier, W. H. & Bounds, W. G. Jr. Reading Statistics and Research. New York: Harper & Row, 1974.

Huff, D. How to Lie with Statistics. New York: Norton, 1954.

Johnson, R. Elementary Statistics. Belmont, CA: Duxbury Press, 1975.

Joiner, B. L. & Campbell, C., "Some Interesting Examples for Teaching Statistics," The Mathematics Teacher, 1975, vol. 68, no. 5, pp. 364-369.

Joiner, B. L. & Campbell, C., "Some Innovations in Statistical Education: With a View Toward a Modular Approach." Modular Instruction in Statistics: Report of the American Statistical Association Study of Modular Instruction, Washington, D. C.: American Statistical Assn., 1976, pp. 84-89.

Keselman, H. J. & Toothaker, L. E., "An Empirical Comparison of the Marascuilo and Normal Scores Non-Parametric Tests, and the Scheffé and Tukey Parametric Tests for Pairwise Contrasts," Proceedings of the 81st Annual American Psychological Association Convention, 1973.

Keselman, H. J., Toothaker, L. E., & Shooter, M., "An Evaluation of Two Unequal N Forms of the Tukey Multiple Comparison Statistic," Journal of the American Statistical Association, 1975, vol. 70, pp. 584-587.

Kirk, R. E. Experimental Design Procedures for the Behavioral Sciences. Monterey, CA: Brooks/Cole, 1969.

Kirk, R. E. Statistical Issues: A Reader for the Behavioral Sciences. Monterey, CA: Brooks/Cole, 1972.

Klausmeier, H., Ghatala, E., & Frayer, D. Conceptual Learning and Development: A Cognitive View. New York: Academic Press, 1974, vol. 1.

Kruskal, W. H. & Tanur, J. M. International Encyclopedia of Statistics. New York: Free Press, 1978, vol. 1 and vol. 2.

Kruskal, W. H., "Significance, Tests of," International Encyclopedia of Statistics. New York: Free Press, 1978, vol. 2, pp. 944-958, (Ed. by Kruskal, W. H. & Tanur, J. M.).

Lefkon, R. H., Fletcher, H. J. & Derderian, J. C., "Statistics Without Calculus for Secondary Through Doctoral Programs," Modular Instruction in Statistics: Report of the American Statistical Association Study of Modular Instruction. (Project Director: Guthrie, D.; Editors: O'Fallon, J. R. & Service, J.), Washington, D. C., 1976, pp. 28-35.

Lieberman, B. Contemporary Problems in Statistics: A Book of Readings for the Behavioral Sciences. New York: Oxford University Press, 1971.

Lindquist, E. F., "The Norton Study of the Effects of Non-Normality and Heterogeneity of Variance," Design of Analysis of Experiments in Psychology and Education. New York: Houghton-Mifflin Co., 1953. (Study first reported in unpublished thesis, "An Empirical Investigation of Some Effects on Non-Normality and Heterogeneity on the F-Distribution,: Norton, D. W., University of Iowa, 1952.)

Linton, M. & Gallo, P. S. Jr. The Practical Statistician: Simplified Handbook of Statistics. Monterey, CA: Brooks/Cole Publishing Co., 1975.

Markle, S. M. & Tiemann, P. W. Really Understanding Concepts: Or in Frumious Pursuit of the Jabberwock. IL: Stipes, 1969.

Mendenhall, W. D. & Ott, L. Understanding Statistics. New York: Duxbury, 1972, (1st ed.).

Mendenhall, W. D. & Ott, L. Understanding Statistics. New York: Duxbury, 1976, (2nd ed.).

Nicewander, W. A. & Price, J. M., "Dependent Variable Reliability and the Power of Significance Tests," Psychological Bulletin, March 1978, pp. 405-508.

Norton, D. W., An Empirical Investigation of Some Effects of Non-Normality and Heterogeneity on the F-Distribution. 1952, University of Iowa, (thesis).

O'Brien, T. C. & Shapiro, B. J., "Statistical Significance - What?", Mathematics Teacher, National Council of Teachers of Mathematics, 1968, vol. 61, pp. 673-676.

Pearson, E. S., "The Analysis of Variance in Cases of Non-Normal Variations," Biometrika, 1931, vol. 23, pp. 114-133.

Pearson, E. S. & Hartley, H. O., "Charts of the Power Function for Analysis of Variance Tests, Derived from the Non-Central F-Distribution," Biometrika, 1951, vol. 38, pp. 112-130.

Poizner, S. B., Alternative Responding and Scoring Methods for Multiple Choice Tests: An Empirical Study. Master's thesis, University of Oklahoma, 1974.

Price, J. M., Program Grader, Developed at the University of Oklahoma, Department of Psychology, 1976. (Available from J. M. Price, Oklahoma State University, Department of Psychology, Stillwater, OK.)

Ramseyer, G. C. & Tcheng, T., "The Robustness of the Studentized Range Statistic to Violations of the Normality and Homogeneity of Variance Assumptions," American Educational Research Journal, 1973, vol. 10, pp. 235-240.

Rosenthal, R. & Gaito, J., "Further Evidence for the Cliff Effects in the Interpretation of Levels of Significance," Psychological Reports, 1964, vol. 15, p. 570.

Rosenthal, R. & Gaito, J., "The Interpretation of Levels of Significance by Psychological Researchers," The Journal of Psychology, 1963, vol. 55, pp. 33-38.

Scheffé, H. The Analysis of Variance. New York: Wiley Press, 1959.

Scheffé, H., "Some Practical Solutions to the Behren-Fisher Problem," Journal of American Statistical Association, 1970, vol. 65, no. 332, pp. 1501-1508.

Schulte, A., "A Case for Statistics," Arithmetic Teacher, 1979, vol. 26, no. 6, p. 24.

Shumway, R. J., "Negative Instances and Mathematical Concept Formation: A Preliminary Study," Journal of Research and Mathematics Education, 1971, vol. 1, no. 3, pp. 218-227.

Shumway, R. J., "Negative Instances in Mathematical Concept Acquisition: Transfer Effects Between the Concepts of Commutativity and Associativity," For Research in Mathematics Education, 1974, vol. 1, no. 3, pp. 218-227.

Smoke, K. L., "Negative Instances in Concept Learning," Journal of Experimental Psychology, 1933, vol. 16, pp. 583-588.

Srivastava, A. B. L., "Effect of Non-Normality on the Power Function of the t-Test," Biometrika, 1958, vol. 45, pp. 421-430.

Steen, L. A. & Seebach, J. A. Counterexamples in Topology. New York: Holt, Rinehart, and Winston, 1971.

Swanson, J. E. The Effects of Number of Positive and Negative Instances, Concept Definition, and Emphasis of Relevant Attributes on the Attainment of Three Environmental Concepts by Sixth-Grade Children. Madison: Wisconsin Research and Development Center for Cognitive Learning, Tech Report, no. 244, 1972.

Tennyson, R. D., Wooley, F. R., & Merrill, M. D., "Exemplar and Non-Exemplar Variables Which Produce Correct Concept Classification Errors," Journal of Educational Psychology, 1972, vol. 63, pp. 144-152.

Toothaker, L. E., "Empirical Investigation of the Permutation t-Test, British Journal of Mathematical and Statistical Psychology, 1972, vol. 25, pp. 83-94.

Toothaker, L. E. Statistics Notes by Toothaker. University of Oklahoma, 1979, (2nd ed.).

Veldman, D. J. Fortran Programming for the Behavioral Sciences. New York: Holt, Rinehart, and Winston, 1967, pp. 129-131.

Welch, B. L., "The Significance of the Difference Between Two Means When the Population Variances are Unequal," Biometrika, 1937, vol. 29, pp. 350-362.

Welch, B. L., "Further Notes on Mrs. Aspin's Tables and on Certain Approximations to the Tabled Functions," Biometrika, 1949, vol. 36, pp. 293-296.

Welch, B. L., "The Generalization of Student's Problem When Several Different Population Variances Are Involved," Biometrika, 1947, vol. 34, pp. 28-35.

Winer, B. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1971, (2nd ed.).

Yates, F., "The Analysis of Multiple Classifications With Unequal Numbers in the Different Classes," Journal of the American Statistical Association, 1934, vol. 29, pp. 51-56.

APPENDIX A

EXPERIMENTAL TREATMENTS

On the first page which follows the reader will see a reduced Xerox copy of the first page of Treatment I (two thirds negative instances, one third positive instances) with the tabs in place as the subjects in Treatment I first saw them. On the second page which follows the reader will see a reduced Xerox copy of that same first page of Treatment I, but with the tabs removed and the correct responses exposed. Each of the following pages are displayed without the tabs and the correct responses exposed.

The order of treatments is: first, Treatment I (two thirds negative instances, one third positive instances); second, Treatment II (one third negative instances, two thirds positive instances); and lastly, Treatment III (all positive instances).

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we have a 2 tailed 2 independent sample t test which is highly insensitive to differences in population means (μ's), i.e. a low power test. If we accept H_0: $\mu_1 = \mu_2$ at the .05 level, with $\bar{X}_1 > \bar{X}_2$, this suggests that:

1. It is likely that μ_1 is very near μ_2. A D
2. It is likely that $\mu_1 = \mu_2$. A D
3. μ_1 may be near, or may be far from μ_2. A D
4. It is likely that $\mu_1 - \mu_2$ is considerably larger than 0. A D
5. It must be that $\mu_1 > \mu_2$. A D
6. It is best to withhold judgment concerning the relationship between μ_1 and μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we test H_0: $\mu_1 = \mu_2$ at the .05 level, with a 2 tailed 2 independent sample t; 4 observations in each group, extremely large variances, and $\bar{X}_1 > \bar{X}_2$. If we accept H_0: $\mu_1 = \mu_2$, this suggests that:

1. It is best to withhold judgment concerning the relationship between μ_1 and μ_2. A D
2. It must be that $\mu_1 > \mu_2$. A D
3. It is likely that $\mu_1 - \mu_2$ is considerably larger than 0. A D
4. μ_1 may be near, or may be far from μ_2. A D
5. It is likely that $\mu_1 = \mu_2$. A D
6. It is likely that μ_1 is very near μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we have a 2 tailed 2 independent sample t test which is highly insensitive to differences in population means (μ's), i.e. a low power test. If we accept H_0: $\mu_1 = \mu_2$ at the .05 level, with $\bar{X}_1 > \bar{X}_2$, this suggests that:

D 1. It is likely that μ_1 is very near μ_2. A D

D 2. It is likely that $\mu_1 = \mu_2$. A D

A 3. μ_1 may be near, or may be far from μ_2. A D

D 4. It is likely that $\mu_1 - \mu_2$ is considerably larger than 0. A D

D 5. It must be that $\mu_1 > \mu_2$. A D

A 6. It is best to withhold judgment concerning the relationship between μ_1 and μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we test H_0: $\mu_1 = \mu_2$ at the .05 level, with a 2 tailed 2 independent sample t; 4 observations in each group, extremely large variances, and $\bar{X}_1 > \bar{X}_2$. If we accept H_0:$\mu_1 = \mu_2$, this suggests that:

A 1. It is best to withhold judgment concerning the relationship between μ_1 and μ_2. A D

D 2. It must be that $\mu_1 > \mu_2$. A D

D 3. It is likely that $\mu_1 - \mu_2$ is considerably larger than 0. A D

A 4. μ_1 may be near, or may be far from μ_2. A D

D 5. It is likely that $\mu_1 = \mu_2$. A D

D 6. It is likely that μ_1 is very near μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose we have a 2 tailed 2 independent sample t test which is highly sensitive to differences in population means (μ's), i.e., high power test. If we accept H_0: $\mu_1 = \mu_2$ at the .05 level, and $\bar{X}_1 > \bar{X}_2$, this indicates that:

D 1. It is certain that $\mu_1 = \mu_2$. A D

A 2. It is very likely that μ_1 is near μ_2. A D

D 3. It is certain that μ_1 is near μ_2. A D

D 4. It is certain that $\mu_1 - \mu_2$ is near 0. A D

D 5. It is likely that $\mu_1 = \mu_2$. A D

A 6. The difference between μ_1 and μ_2 is likely to be small. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we accept H_0: $\mu_1 = \mu_2$ at the .05 level, in a 2 tailed 2 independent sample t test; n = 40,000 observations in each group, very small S^2s, and $\bar{X}_1 > \bar{X}_2$, this indicates that:

A 1. The difference between μ_1 and μ_2 is likely to be small. A D

D 2. It is likely that $\mu_1 = \mu_2$. A D

D 3. It is certain that $\mu_1 - \mu_2$ is near 0.

D 4. It is certain that μ_1 is near μ_2. A D

A 5. It is very likely that μ_1 is near μ_2. A D

D 6. It is certain that $\mu_1 = \mu_2$. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose we have a statistical test which is highly sensitive to the differences in population means (i.e. a high power test). If we reject H_0: $\mu_1 = \mu_2$ at the .05 level, with $\bar{X}_1 > \bar{X}_2$, then this suggests that:

D 1. It is unlikely that μ_1 is near μ_2. A D

D 2. It must be that μ_1 is nowhere near μ_2. A D

A 3. $\mu_1 - \mu_2$ may be small, or it may be large, but it is likely that $\mu_1 - \mu_2 > 0$. A D

D 4. It is likely that μ_1 is much larger than μ_2. A D

D 5. It is definite that $\mu_1 > \mu_2$. A D

A 6. This sensitive test might reject $\mu_1 = \mu_2$, even when μ_1 and μ_2 are close. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we test H_0: $\mu_1 = \mu_2$ at the .05 level, with a 2 tailed 2 independent sample t; 50,000 observations in each group very small sample variances, and $\bar{X}_1 > \bar{X}_2$. If we reject H_0: $\mu_1 = \mu_2$, then this suggests that:

A 1. This sensitive test might reject $\mu_1 = \mu_2$, even when μ_1 and μ_2 are close. A D

D 2. It is definite that $\mu_1 > \mu_2$. A D

D 3. It is likely that μ_1 is much larger than μ_2. A D

A 4. $\mu_1 - \mu_2$ may be small, or it may be large, but it is likely that $\mu_1 - \mu_2 > 0$. A D

D 5. It must be that μ_1 is nowhere near μ_2. A D

D 6. It is unlikely that μ_1 is near μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we reject H_0: $\mu_1 = \mu_2$ in a very low power test, that is, very insensitive to differences in population means; $\bar{X}_1 > \bar{X}_2$, this indicates:

D 1. It is certain that $\mu_1 > \mu_2$. A D

A 2. It is very likely that μ_1 is considerably larger than μ_2. A D

D 3. It is certain that μ_1 is considerably larger than μ_2. A D

D 4. It is certain that $\mu_1 - \mu_2$ is large. A D

D 5. It is certain that $\mu_1 \neq \mu_2$. A D

A 6. The difference between μ_1 and μ_2 is likely to be fairly large. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we reject H_0: $\mu_1 = \mu_2$ at the .05 level, with $n = 4$ observations in each group; very large S^2s and $\bar{X}_1 > \bar{X}_2$, this indicates that:

A 1. The difference between μ_1 and μ_2 is likely to be fairly large. A D

D 2. It is certain that $\mu_1 \neq \mu_2$. A D

D 3. It is certain that $\mu_1 - \mu_2$ is large. A D

D 4. It is certain that μ_1 is considerably larger than μ_2. A D

A 5. It is very likely that μ_1 is considerably larger than μ_2 A D

D 6. It is certain that $\mu_1 > \mu_2$. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we have a 2 tailed 2 independent sample t test which is highly insensitive to differences in population means (μ's), that is a low power test. If we accept $H_0: \mu_1 = \mu_2$ at the .05 level, with $\bar{X}_1 > \bar{X}_2$, this suggests that:

D 1. It is likely that μ_1 is very near μ_2. A D

A 2. μ_1 may be near, or may be far from μ_2. A D

A 3. $\mu_1 - \mu_2$ may be small, or it may be large. A D

D 4. It must be that $\mu_1 > \mu_2$. A D

A 5. Since the test is highly insensitive to unequal μ's, we may have failed to detect a meaningful difference between μ's. A D

A 6. It is best to withhold judgment concerning the relationship between μ_1 and μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we test $H_0: \mu_1 = \mu_2$ at the .05 level, with a 2 tailed 2 independent sample t; 4 observations in each group, extremely large variances, and $\bar{X}_1 > \bar{X}_2$. If we accept $H_0: \mu_1 = \mu_2$, this suggests that:

A 1. It is best to withhold judgment concerning the relationship between μ_1 and μ_2. A D

A 2. Since the test is highly insensitive to unequal μ's, we may have failed to detect a meaningful difference between μ's, A D

D 3. It must be that $\mu_1 > \mu_2$. A D

A 4. $\mu_1 - \mu_2$ may be small, or it may be large. A D

A 5. μ_1 may be near, or it may be far from μ_2. A D

D 6. It is likely that μ_1 is very near μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose we have a 2 tailed 2 independent sample t test which is highly sensitive to differences in population means (μ's), i.e., high power test. If we accept H_0: $\mu_1 = \mu_2$ at the .05 level, and $\bar{X}_1 > \bar{X}_2$, this indicates that:

D 1. It is certain that $\mu_1 = \mu_2$. A D

A 2. It is very likely that μ_1 is near μ_2. A D

A 3. It is very likely that μ_1 and μ_2 are close. A D

A 4. It is very unlikely that $\mu_1 - \mu_2$ is large. A D

D 5. It is certain that μ_1 is near μ_2. A D

A 6. The difference between μ_1 and μ_2 is likely to be small. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we accept H_0: $\mu_1 = \mu_2$ at the .05 level, in a 2 tailed 2 independent sample t test; n =40,000 observations in each group, very small S^2s, and $\bar{X}_1 > \bar{X}_2$, this indicates that:

A 1. The difference between μ_1 and μ_2 is likely to be small. A D

D 2. It is certain that μ_1 is near μ_2. A D

A 3. It is very unlikely that $\mu_1 - \mu_2$ is large. A D

A 4. It is very likely that μ_1 and μ_2 are close. A D

A 5. It is very likely that μ_1 is very near μ_2. A D

D 6. It is certain that $\mu_1 = \mu_2$. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose we have a statistical test which is highly sensitive to the differences in population means (i.e a high power test). If we reject H_0: $\mu_1 = \mu_2$ at the .05 level. with $\bar{X}_1 > \bar{X}_2$, then this suggests that:

D 1. It is unlikely that μ_1 is near μ_2. A D

A 2. $\mu_1 - \mu_2$ may be small, or may be large, but it is likely that $\mu_1 - \mu_2 > 0$. A D

A 3. This sensitive test may have rejected H_0: $\mu_1 = \mu_2$ due to a small difference between means A D

D 4. It is likely that μ_1 is much larger than μ_2. A D

A 5. It is likely that $\mu_1 > \mu_2$, but we cannot say how much greater. A D

A 6. This sensitive test might reject $\mu_1 = \mu_2$, even when μ_1 and μ_2 are close. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we test H_0: $\mu_1 = \mu_2$ at the .05 level, with a 2 tailed 2 independent sample t; 40,000 observations in each group, very small sample variances, and $\bar{X}_1 > \bar{X}_2$. If we **reject** H_0: $\mu_1 = \mu_2$, this suggests:

A 1. This sensitive test might reject $\mu_1 = \mu_2$, even when μ_1 and μ_2 are close. A D

A 2. It is likely that $\mu_1 > \mu_2$, but we cannot say how much greater. A D

D 3. It is likely that μ_1 is much larger than μ_2. A D

A 4. This sensitive test may have rejected H_0: $\mu_1 = \mu_2$ due to a small difference between means. A D

A 5. $\mu_1 - \mu_2$ may be small, or may be large, but it is likely that $\mu_1 - \mu_2 > 0$. A D

D 6. It is unlikely that μ_1 is near μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we reject H_0: $\mu_1 = \mu_2$ in a very low power test, that is, very insensitive to differences in population means: $\bar{X}_1 > \bar{X}_2$, this indicates:

A 1. It is very likely that μ_1 is larger than μ_2. A D

D 2. It is certain that μ_1 is considerably larger than μ_2. A D

A 3. It is very likely that μ_1 is considerably larger than μ_2. A D

D 4. It is certain that $\mu_1 - \mu_2$ is large. A D

A 5. It is very unlikely that μ_1 and μ_2 are extremely close. A D

A 6. The difference between μ_1 and μ_2 is likely to be fairly large. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we reject H_0: $\mu_1 = \mu_2$ at the .05 level, with $n = 4$ observations in each group; very large S^2s, and $\bar{X}_1 > \bar{X}_2$, this indicates that:

A 1. The difference between μ_1 and μ_2 is likely to be fairly large. A D

A 2. It is very unlikely that μ_1 and μ_2 are extremely close. A D

D 3. It is certain that $\mu_1 - \mu_2$ is large. A D

A 4. It is very likely that μ_1 is considerably larger than μ_2. A D

D 5. It is certain that μ_1 is considerably larger than μ_2. A D

A 6. It is very likely that μ_1 is larger than μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we have a 2 tailed 2 independent sample t test which is highly insensitive to differences in population means (μ's), that is, a low power test. If we accept $H_0: \mu_1 = \mu_2$ at the .05 level, with $\bar{X}_1 > \bar{X}_2$, this suggests that:

A 1. μ_1 may be near, or may be far from μ_2. A D

A 2. $\mu_1 - \mu_2$ may be small, or it may be large. A D

A 3. Since this test is highly insensitive to unequal μ's, we may have failed to detect a meaningful difference between μ's. A D

A 4. We would likely accept $H_0: \mu_1 = \mu_2$, even if μ_1 is moderately far from μ_2. A D

A 5. $\mu_1 - \mu_2$ may be near, or may be far from 0. A D

A 6. It is best to withhold judgment concerning the relationship between μ_1 and μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we test $H_0: \mu_1 = \mu_2$ at the .05 level, with a 2 tailed 2-independent sample t; 4 observations in each group, extremely large variances, and $\bar{X}_1 > \bar{X}_2$. If we accept $H_0: \mu_1 = \mu_2$, this suggests that:

A 1. It is best to withhold judgment concerning the relationship between μ_1 and μ_2. A D

A 2. $\mu_1 - \mu_2$ may be near, or may be far from 0. A D

A 3. We would likely accept $H_0: \mu_1 = \mu_2$, even if μ_1 is moderately far from μ_2. A D

A 4. Since this test is highly insensitive to unequal μ's, we may have failed to detect a meaningful difference between μ's. A D

A 5. $\mu_1 - \mu_2$ may be small, or it may be large. A D

A 6. μ_1 may be near, or may be far from μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose we have a 2 tailed 2 independent sample t test which is highly sensitive to differences in population means (μ's). I.e., high power test.

If we accept $H_0: \mu_1 = \mu_2$ at the .05 level, and $\bar{X}_1 > \bar{X}_2$, this indicates that:

A 1. It is very likely that μ_1 is near μ_2. A D

A 2. It is very likely that $\mu_1 - \mu_2$ is near 0. A D

A 3. It is very likely that μ_1 and μ_2 are close. A D

A 4. It is very unlikely that μ_1 and μ_2 are far apart. A D

A 5. It is very unlikely that $\mu_1 - \mu_2$ is large. A D

A 6. The difference between μ_1 and μ_2 is likely to be small. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we accept $H_0: \mu_1 = \mu_2$ at the .05 level, in a 2 tailed 2 independent sample t test; n = 40,000 observations in each group, very small S^2s, and $\bar{X}_1 > \bar{X}_2$, this indicates that:

A 1. The difference between μ_1 and μ_2 is likely to be small. A D

A 2. It is very unlikely that $\mu_1 - \mu_2$ is large. A D

A 3. It is very unlikely that μ_1 and μ_2 are far apart. A D

A 4. It is very likely that μ_1 and μ_2 are close. A D

A 5. It is very likely that $\mu_1 - \mu_2$ is near 0. A D

A 6. It is very likely that μ_1 is near μ_2. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose we have a statistical test which is highly sensitive to differences in population means (i.e. a high power test). If we reject $H_0: \mu_1 = \mu_2$ at the .05 level, with $\bar{X}_1 > \bar{X}_2$, then this suggests that:

A 1. μ_1 may be near, or may be far from μ_2, but it is likely that $\mu_1 > \mu_2$. A D

A 2. $\mu_1 - \mu_2$ may be small, or may be large, but it is likely that $\mu_1 - \mu_2 > 0$. A D

A 3. If $\mu_1 = \mu_2$, an outcome at least this extreme could occur with probability less than .05. A D

A 4. This sensitive test may have rejected $H_0: \mu_1 = \mu_2$ due to a small difference between means. A D

A 5. It is likely that $\mu_1 > \mu_2$, but we cannot say how much greater. A D

A 6. This sensitive test might reject $\mu_1 = \mu_2$, even when μ_1 and μ_2 are close. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

Suppose that we test $H_0: \mu_1 = \mu_2$ at the .05 level, with a 2 tailed 2 independent sample t; 40000 observations in each group, very small sample variances, and $\bar{X}_1 > \bar{X}_2$. If we reject $H_0: \mu_1 = \mu_2$, this suggests that:

A 1. This sensitive test might reject $\mu_1 = \mu_2$, even when μ_1 and μ_2 are close. A D

A 2. It is likely that $\mu_1 > \mu_2$, but we cannot say how much greater. A D

A 3. This sensitive test may have rejected $H_0: \mu_1 = \mu_2$ due to a small difference between means. A D

A 4. If $\mu_1 = \mu_2$, an outcome at least this extreme could occur with probability less than .05. A D

A 5. $\mu_1 - \mu_2$ may be small, or may be large, but it is likely that $\mu_1 - \mu_2 > 0$. A D

A 6. μ_1 may be near, or may be far from μ_2, but it is likely that $\mu_1 > \mu_2$. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we reject $H_0: \mu_1 = \mu_2$ in a very low power test, that is, very insensitive to differences in population means: $\bar{X}_1 > \bar{X}_2$, this indicates:

A 1. It is very likely that μ_1 is larger than μ_2. A D

A 2. It is very likely that μ_1 is considerably larger than μ_2. A D

A 3. It is very likely that $\mu_1 - \mu_2$ is fairly large. A D

A 4. It is very unlikely that μ_1 and μ_2 are extremely close. A D

A 5. It is very unlikely that $\mu_1 - \mu_2$ is very small. A D

A 6. The difference between μ_1 and μ_2 is likely to be fairly large. A D

ASSUME ALL ASSUMPTIONS ARE MET FOR THE 2-INDEPENDENT SAMPLE t.

If we reject $H_0: \mu_1 = \mu_2$ at the .05 level, with $n = 4$ observations in each group; very large S^2s, and $\bar{X}_1 > \bar{X}_2$, this indicates that

A 1. The difference between μ_1 and μ_2 is likely to be fairly large. A D

A 2. It is very unlikely that $\mu_1 - \mu_2$ is very small. A D

A 3. It is very unlikely that μ_1 and μ_2 are extremely close. A D

A 4. It is very likely that $\mu_1 - \mu_2$ is fairly large. A D

A 5. It is very likely that μ_1 is considerably larger than μ_2. A D

A 6. It is very likely that μ_1 is larger than μ_2. A D

APPENDIX B

CRITERION TEST

FOR EACH ALTERNATIVE OF EACH ITEM, GIVE A RANK OF "3" to "0"; "3" BEST RESPONSE, "2" NEXT CHOICE, ETC. EVALUATE EACH ITEM CAREFULLY BEFORE CHOOSING.

ASSUME ALL UNDERLYING ASSUMPTIONS ARE MET FOR THE 2 tailed 2 independent SAMPLE t TEST WHICH IS BEING USED IN QUESTIONS 1—15.

Each of 2 groups containing n rats is fed one of two diets for 6 weeks. It is then desired to test H_0: $\mu_1 = \mu_2$ where μ_1 is the population mean weight for rats fed diet 1, and μ_2 is the population mean weight for rats fed diet 2.

1. If H_0: $\mu_1 = \mu_2$ is rejected at the .05 level, when $\bar{X}_1 > \bar{X}_2$ in a very high power test, then it is __ true that μ_1 is much greater than μ_2:

___a) certainly

___b) likely

___c) neither likely, nor unlikely

___d) unlikely

2. If H_0: $\mu = \mu_2$ is rejected at the .05 level, when $\bar{X}_1 > \bar{X}_2$ in a very low power test, then it is __ true that μ_1 is much larger than μ_2:

___ a) certainly

___ b) likely

___ c) neither likely, nor unlikely

___ d) unlikely

3. If H_0: $\mu_1 = \mu_2$ is rejected at the .05 level, when $\bar{X}_1 > \bar{X}_2$ in a very low power test, then it is __ true that μ_1 is very near μ_2:

___ a) certainly

___ b) likely

___ c) neither likely, nor unlikely

___d) unlikely

4. If H_0: $\mu_1 = \mu_2$ is rejected at the .05 level, when $\bar{X}_1 > \bar{X}_2$, n = 40,000, and very small S^2c, then it is __ true that μ_1 is much larger than μ_2:

___ a) certainly

___ b) likely

___ c) neither likely, nor unlikely

___ d) unlikely

5. If H_0: $\mu_1 = \mu_2$ is rejected at the .05 level when $\bar{X}_1 > \bar{X}_2$, $n = 4$, and s^2's are fairly large, then it is __ true that μ_1 is much larger than μ_2:

__a) certainly

__b) likely

__c) neither likely, nor unlikely

__d) unlikely

6. If H_0: $\mu_1 = \mu_2$ is rejected at the .05 level, then it is __ true that if $\mu_1 = \mu_2$, the probability of obtaining a t value this extreme, or more, is .05:

__a) certainly

__b) likely

__c) neither likely, nor unlikely

__d) unlikely

7. If H_0: $\mu_1 = \mu_2$ is accepted at the .05 level, when $\bar{X}_1 > \bar{X}_2$ in a very high power test, then it is __ true that μ_1 is much larger than μ_2:

__a) certainly

__b) likely

__c) neither likely, nor unlikely

__d) unlikely

8. If H_0: $\mu_1 = \mu_2$ is accepted at the .05 level when $\bar{X}_1 > \bar{X}_2$ in a very low power test, then it is __ true that μ_1 is very near μ_2:

__a) certainly

__b) likely

__c) neither likely, nor unlikely

__d) unlikely

9. If H_0: $\mu_1 = \mu_2$ is accepted at the .05 level, with $\bar{X}_1 > \bar{X}_2$ in a very high power test, then it is __ true that μ_1 is very near μ_2:

__a) certainly

__b) likely

__c) neither likely, nor unlikely

__d) unlikely

10. If H_0: $\mu_1 = \mu_2$ is accepted at the .05 level when $\bar{X}_1 > \bar{X}_2$, n = 3, and s^2s are very large, then it it __ true that μ_1 is very near μ_2:

___a) certainly

___b) likely

___c) neither likely, nor unlikely

___d) unlikely

11. If H_0: $\mu_1 = \mu_2$ is accepted at the .05 level when $\bar{X}_1 > \bar{X}_2$, n = 40,000, and very small s^2s, then it is __ true that μ_1 is very near μ_2:

___a) certainly

___b) likely

___c) neither likely, nor unlikely ___d) unlikely

12. If H_0: $\mu_1 = \mu_2$ is accepted at the .05 level when $\bar{X}_1 > \bar{X}_2$, n = 40,000, and very small s^2s, then it is __ true that μ_1 is much larger than μ_2:

___a) certainly

___b) likely

___c) neither likely, nor unlikely

___d) unlikely

13. If we reject H_0: $\mu_1 = \mu_2$ at the .05 level, and $\bar{X}_1 > \bar{X}_2$, then it is __ true that $\mu_1 > \mu_2$:

___a) certainly

___b) likely

___c) neither likely, nor unlikely

___d) unlikely

14. When we learn that an investigator has rejected H_0: $\mu_1 = \mu_2$ at the .05 level with $\bar{X}_1 > \bar{X}_2$, we can be absolutely certain that:

___a) $\mu_1 > \mu_2$. ___b) μ_1 is nowhere near μ_2.

___c) if $\mu_1 = \mu_2$, sample data this extreme, or more so, could occur less than 5% of the time, on the average.

___d) it is unlikely that μ_1 is near μ_2.

15. When it is learned that an investigator has accepted H_0: $\mu_1 = \mu_2$ at the .05 level, with $\bar{X}_1 > \bar{X}_2$, we can be absolutely certain:

___a) $\mu_1 \geq \mu_2$. ___b) μ_1 is near μ_2.

___c) it is likely that μ_1 is near μ_2.

___d) only that H_0: $\mu_1 = \mu_2$ was not rejected.

APPENDIX C

SOME ITEMS FROM A QUIZ GIVEN THE WEEK

PRIOR TO THE EXPERIMENT

Pts.

FOR EACH ALTERNATIVE OF EACH ITEM, GIVE A RANK OF "3" to "0"; "3" BEST RESPONSE, "2" FOR NEXT CHOICE, ETC. EVALUATE EACH ITEM BEFORE CHOOSING.

1. If H_o: $\mu = 9$, $\alpha = .05$, and power = .85 for H_1: $\mu = 9.25$, then:

___ a) Probability we reject H_o when $\mu = 9$ is .05.

___ b) Probability we reject H_o when $\mu = 9.25$ is .85.

___ c) Probability we accept H_o when $\mu = 9.25$ is .15.

___ d) All of the above are true.

2. In the hypothesis tests which we are considering, when $\alpha = .05$, for power to increase, we would want:

___ a) sample size to increase, and variance decrease.

___ b) sample size to decrease, and variance increase.

___ c) sample size and variance both to increase.

___ d) sample size and variance both to decrease.

3. Power can be thought of intuitively as a measure of:

___ a) insensitivity of the test to deviations from H_o.

___ b) sensitivity of the test to deviations from H_o.

___ c) insensitivity of the test to violations of assumptions.

___ d) 1 - P (Type I error).

4. When H_o: $\mu = 0$, then if in reality $\mu = 1$, the hypothesis test we use would be most likely to reject H_o if for H_1: $\mu = 1$, the test had:

___ a) low power.

___ b) moderate to high power.

___ c) very high power.

___ d) We would always reject H_o: $\mu = 0$ when $\mu = 1$, regardless of power.

APPENDIX D

RAW SCORE DATA

	Group 1	Group 2	Group 3	Group 4
ABOVE	45	44	45	42
MEDIAN	44	43	43	41
	43	42	42	36
	43	40	41	32
	41	39	40	30
	40	38	35	27
	39	35	33	
BELOW	41	41	36	40
MEDIAN	41	40	40	34
	39	35	34	31
	38	32	31	26
	35	31	28	25
	35	26	28	25
	29	26	23	22

APPENDIX E

SUMMARY STATISTICS FOR TREATMENT BLOCK CELLS

	Two Thirds Negative	One Third Negative	All Positive	Control
ABOVE MEDIAN	$\bar{X} = 42.14$ $S = 2.19$	$\bar{X} = 40.14$ $S = 3.13$	$\bar{X} = 39.85$ $S = 4.34$	$\bar{X} = 34.66$ $S = 6.06$
BELOW MEDIAN	$\bar{X} = 36.86$ $S = 4.26$	$\bar{X} = 33.00$ $S = 6.06$	$\bar{X} = 31.42$ $S = 5.71$	$\bar{X} = 29.00$ $S = 6.32$

APPENDIX F

SUMMARY STATISTICS FOR EACH OF THE FOUR TREATMENTS

Two Thirds Negative	One Third Negative	All Positive	Control
$\bar{X}$ = 39.50	$\bar{X}$ = 36.57	$\bar{X}$ = 35.64	$\bar{X}$ = 31.62
S = 4.26	S = 5.93	S = 6.55	S = 6.63

APPENDIX G

RESULTS OF PAIRWISE TUKEY-TESTS

Contrast K-K'	Aspin-Welch WSD (obs. $q = \sqrt{2}v$)	d.f.(Aspin-Welch)	[Aspin-Welch] d.f
1-2	2.122571257	23.581122037	23
1-3	2.613714059	22.323245336	22
1-4	5.158384203*	20.214624684	20
2-3	.556172368	25.750940328	25
2-4	2.887468333	24.155155617	24
3-4	2.244367773	24.802191764	24

*p < .05

APPENDIX H

ANALYSIS

The consequences of violation of assumptions for the one way ANOVA are as indicated in Chapter 10 of Scheffé (1959), in the review article by Glass et al. (1972), research by Toothaker (1972), Toothaker and Feir (1974), and in an article by Elashoff and Elashoff, in the International Encyclopedia of Statistics (Kruskal & Tanur, 1978). To very briefly summarize these results:

When alpha is set equal to .05, then moderate deviations from normality will not severely distort the true alpha value; i.e., it is likely that .025 < alpha true < .08. When sample sizes are equal, and N > 10, then unequal population variances, even including as large as ten to one ratios; will not seriously distort the true value of alpha from .05 for the two independent sample case. For K independent sample case, alpha true > .05 to a slight degree. When sample sizes are unequal, then the inequality of population variances can seriously distort both the true alpha values and power. As dispersion of sample sizes, or population variances increase, the degree of distortion becomes more serious.

In the present study, the sample sizes were 14 for each of the experimental groups, and 13 for the control group. From inspection of the sample variances, the assumption of equal population variances does not seem tenable to the author. In fact, it appears that the group having the smaller sample size would have the largest variance. Based on studies reported in the above cited references, the author feels that the true P value for the one way ANOVA on treatments, while likely larger than the value reported (.0112), would not be much larger than .05.

More convincing evidence for mean differences, the author feels, can be found in the experimentwise Aspin-Welch modification of the Tukey Test for pairwise mean differences. This Aspin-Welch modification of the Tukey Test is relatively insensitive to unequal variances in the presence of unequal sample sizes (Games & Howell, 1976). The test statistic associated with the Aspin-Welch modification of the Tukey Test, when multiplied by the square root of 2, is distributed under the null hypothesis when assumptions are met, according to the studentized range. Studentized range statistics have been shown to be relatively insensitive to non-normality when sample sizes are equal (Ramseyer & Tcheng, 1973).

Since the sample sizes were approximately equal, and since the pairwise mean difference between the control group and the group exposed to the two thirds negative instances was significant at the .05 experimentwise error rate, using the Aspin-Welch Tukey Test, the case for a true difference between these two population means seems tenable.

APPENDIX I

ASPIN-WELCH VERSION OF TUKEY'S WSD TEST

In order to determine whether any of the pairwise mean differences were significant, a modification of Tukey's Wholly Significant Difference (WSD) Test was employed in this study. The above modification uses the Aspin-Welch statistic v with approximate degrees of freedom, as developed by B. L. Welch (1949). In this form of Tukey's WSD, H_o: $\mu_j = \mu_{j'}$ is rejected iff $|v| \geq q(\alpha, j, \nu) / \sqrt{2}$ where α equals family-wise, or experimentwise, probability of Type I error, J equals the number of treatments, and ν equals the approximation of Welch degrees of freedom.

In testing the pair-wise mean difference H_o: $\mu_j = \mu_{j'}$, we compute the Aspin-Welch statistic:

$$v = \frac{\bar{Y}_{\cdot j \cdot} - \bar{Y}_{\cdot j' \cdot}}{\sqrt{\frac{S^2_j}{n_j} + \frac{S^2_{j'}}{n_{j'}}}}$$

where S^2_j equals the unbiased sample variance of treatment j; n_j equals the number of observations for treatment j; and where $\bar{Y}_{\cdot j \cdot}$ equals the mean of the sample scores of the treatment j. We reject H_o iff $\sqrt{2}\ |v| \geq q(\alpha, J, \nu)$, where α equals the probability of Type I error,

J equals the number of treatments, $q(\alpha, J, \nu)$ represents the critical value for the studentized range statistic, and ν equals the approximation of Welch degrees of freedom which is computed in the following manner:

$$\nu = \frac{\left(\frac{S^2_j}{n_j} + \frac{S^2_{j'}}{n_{j'}} \right)^2}{\frac{\left(\frac{S_j}{n_j} \right)^2}{n_j - 1} + \frac{\left(\frac{S_{j'}}{n_{j'}} \right)^2}{n_{j'} - 1}}$$

In practice, often the greatest integer function is used for the degrees of freedom, although sometimes interpolation is also used. In the present study, the greatest integer function is used.

The Aspin-Welch statistic was developed specifically for situations in which the author had an unequal number of observations per cell and suspected unequal population variances, and was testing H_o: $\mu_j = \mu_{j'}$. Due to inequality of variances, the author hesitated to use the two independent sample t. Also, Scheffé (1971) has described the Aspin-Welch as the best solution to date for testing H_o: $\mu_1 = \mu_2$ in the presence of unequal variances and unequal sample sizes. Games and Howell (1976) adapted the Aspin-Welch statistic to Tukey's WSD Test and found that it gave relatively good protection, with respect to family-wise Type I and Type II errors in the presence of unequal sample sizes and unequal population variances, particularly when the smallest $n \geq 6$.

Ramseyer and Tcheng (1973) found the studentized range should be relatively insensitive to violations of the normality assumptions when sample sizes are equal. For the sample sizes used in the present study

the protection would probably be less secure, but still, in the author's opinion, reasonably accurate.

Kirk (1969) recommends using the Tukey WSD, but with the harmonic means of the sample size used in place of the common sample size. In research conducted by Toothaker, et al. (JASA, 1975), this procedure was found to be non-robust to unequal variances in the presence of unequal sample sizes. The inequality of the sample sizes in the present study is so slight, that even where the likely large differences occur in population variances, the liberal bias would probably not be large. By using Kirk's modification of the Tukey WSD in the present study, the results obtained were the same as those yielded by the Aspin-Welch modification.

APPENDIX J

POWER

In statistics, the power of a statistical text can be thought of intuitively as a measure of the sensitivity of that test to deviations from the null hypothesis. More formally, for a given hypothesis H, the power of the test P (H) = the probability that we reject H_o when H is true. It follows from the previous statement that:

1. When $H = H_o$, $P(H_o)$ = alpha, the probability of a Type I error.
2. When $H = H_a$, a given alternative hypothesis, the power of the test $P(H_a) = 1 - \beta(H_a)$, where $\beta(H_a)$ = the probability of a Type II error when H_a is true.

It is important for researchers to know the power of a test, whether or not H_o is rejected. In fact, power should guide the researcher in the sample size choice. Conversely, a knowledge of the sample size and sample variance, the value set for alpha, and other factors influencing power can help us interpret more appropriately the meaning of a rejected, or a non-rejected null hypothesis.

Rejection of H_o in a very high power test could possibly be due

to a trivial deviation from H_0. If, however, H_0 is rejected in a very low power test, one would feel more confident that the deviation from H_0 was non-trivial.

If H_0 is accepted in a very high power test, one could feel quite confident that H_0 was approximately true. If, however, H_0 is accepted in a very low power test, one could only say that there was insufficient evidence to reject H_0.

For most statistical tests related to the univariate analysis of variance (ANOVA), power increases as:

sample size increases,

alpha increases,

reliability of measuring instrument increases,

extraneous variability decreases,

deviations from H_0 increase.

When H_0 is true and assumptions are met, the sampling distribution of the F-ratio, in repeated sampling from the same population, is distributed according to the theoretical F-distribution with appropriate degrees of freedom. However, when H_0 is false, in repeated sampling, F-observed is distributed according to a non-central F-distribution whose form depends, not only on the appropriate degrees of freedom, but also on ϕ, a non-centrality parameter. The following is an example for determining power from sample size, using the one-way fixed effects ANOVA completely randomized design, with $\alpha = .05$. For this design, the non-centrality parameter $\phi = \sqrt{\frac{I\sum_{j=1}^{J} \alpha_j^2}{J\sigma^2}}$ where:

I = the number of subjects per treatment;

α_j = the effect due to being in treatment J;

J = the number of treatments;

σ^2 = the common population variance of each treatment.

Suppose we have, as in the present study, a design involving J = 4 treatments (having collapsed over the blocking variable), and 14 subjects in each group, except the control group which has 13 subjects. A lower bound for our power would be the minimum power value obtained from the Pearson-Hartley charts for a one-way fixed effects ANOVA with four levels, and I = 13 scores per group.

We outline how to determine the power of a design having four treatments and 13 observations per treatment, against an alternative maximum difference of $b\sigma$ among means. The value of $\phi = \sqrt{\frac{I\sum_{j=1}^{J} \alpha_j^2}{J\sigma^2}}$ is minimized when $\alpha_1 = \frac{b\sigma}{2}$, $\alpha_2 = \frac{-b\sigma}{2}$, $\alpha_3 = 0$, $\alpha_4 = 0$. Since power is an increasing function of ϕ, this choice of the α_j's yields a minimum value for power.

Should we decide to compute the minimum power for a maximum difference of two standard deviations between any pair of population means for this four level design with 13 scores per group, then J = 4, I = 13, b = 2, and $\frac{b}{2} = 1$.

(In the following formula, note the coefficient "I" appears before the summation, since there are "I" scores in each group.)

Thus, since $\phi = \sqrt{\frac{I\sum_{j=1}^{J} \alpha_j^2}{J\sigma^2}}$ where $\alpha_1 = -\sigma$, $\alpha_2 = \sigma$, $\alpha_3 = 0$, $\alpha_4 = 0$.

Then, $\sum_{j=1}^{J} \alpha_j^2 = 2\sigma^2$, and $\phi = \sqrt{\frac{(13)\ (2\sigma^2)}{(4)\ \sigma^2}} = \sqrt{6.5} = 2.55.$

Entering the Pearson-Hartley chart with $\nu_1 = J - 1 = 3$ d.f., $\nu_2 = IJ - J = 48$ d.f., we find that the minimum power for a maximum difference of two standard deviations among population means is at least .988.

Should we wish to compute the minimum power for a maximum difference of one standard deviation between any pair of population means for the same example, then, since now b = 1, all that must be changed is that $\frac{b}{2} = \frac{1}{2}$. Thus, $\alpha_1 = -\frac{1}{2}\sigma$, $\alpha_2 = \frac{1}{2}\sigma$, $\alpha_3 = 0$, and $\alpha_4 = 0$. Then,

$\sum_{j=1}^{J} \alpha_j^2 = \frac{1}{2}\sigma^2$, and $\phi = \sqrt{\frac{(13)\ (\frac{1}{2}\sigma^2)}{(4)\ \sigma^2}} = \sqrt{1.625} \simeq 1.27.$

The minimum power against a maximum mean difference of one standard deviation is approximately .4.

APPENDIX K

A NON-EXPERIMENTAL CONSIDERATION

An interesting non-experimental aspect associated with the study, involved interviews with former elementary statistics students from both the mathematics department's version of the course, and the psychology department's development of a reasonably similar course. This group of ten people also comprised a convenience sample, and was deliberately biased in favor of C-level students.

In each case, the former student participated in one form of the experimental treatments, took the criterion test, examined the other experimental treatments in which they had not participated, and then discussed their feelings about the study with the author.

All of these former students expressed the opinion that participation in the experiment would be a helpful learning experience for elementary statistics students. It was also generally agreed that the treatments involving negative instances would likely be the most helpful. Some division of opinion arose on the question of whether the two thirds negative instance treatments or the one third would work best. Of those who did express a preference, a majority chose the one third negative,

two thirds positive treatments as being the most effective. One of those students said he felt that a priori probability of circling the "A" would be greater than one half, given that the student was not sure of the correct response. Thus, he concluded, such a student would obtain positive reinforcement more frequently on the one third than on the two thirds negative instance treatments. He was then asked if an extension of the same reasoning would not suggest that all positive instances would be optimal. The reply was that too many positive instances would bore many students, and would not sufficiently focus attention on the limits of the meaning of the concept.

These informal conversations with former students were very helpful to the author in many ways, and may also be useful to the reader, in that the preference for one third negative instance treatments, as expressed by some of these former students, when considered with the above stated limitations, suggests that the empirical superiority of the two thirds negative instances, suggested by the results of the experiment, should be interpreted very tentatively (Aspin-Welch). The difference between the two thirds negative instances and the one third negative instances did not approach significance.