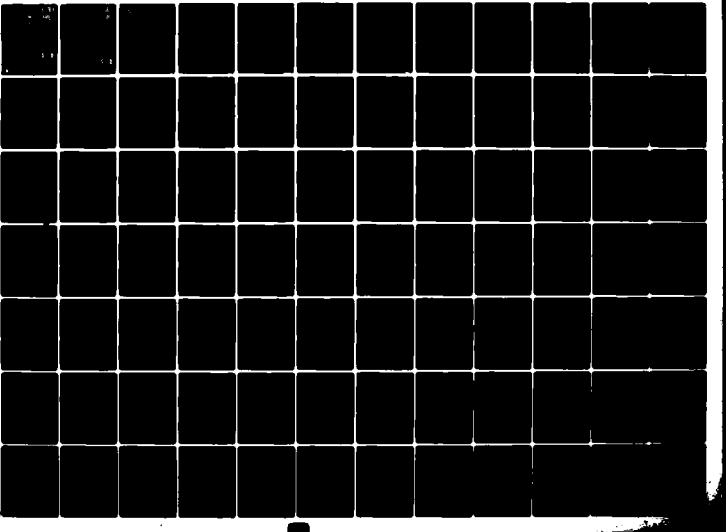
AD-A093 202 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL--ETC F/G 6/16
AN INVESTIGATION OF EXPERIMENT DESIGNS FOR APPLICATIONS IN BIOF--ETC(U)
SEP 80 G FRIED
UNCLASSIFIED AFIT/LSSR-83-80 NL

1 OF 2
AD A093 202



DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY (ATC)

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

3

LEVEL II

AN INVESTIGATION OF EXPERIMENT DESIGNS FOR APPLICATIONS IN BIOFEEDBACK-PERFORMANCE RESEARCH METHODOLOGIES

Gilbert Fried, Captain, USAF

LSSR 83-80

The contents of the document are technically accurate, and no sensitive items, detrimental ideas, or deleterious information are contained therein. Furthermore, the views expressed in the document are those of the author(s) and do not necessarily reflect the views of the School of Systems and Logistics, the Air University, the Air Training Command, the United States Air Force, or the Department of Defense.

USAF SCN 75-20B AFIT Control Number LSSR 83-80

AFIT RESEARCH ASSESSMENT

The purpose of this questionnaire is to determine the potential for current and future applications of AFIT thesis research. Please return completed questionnaires to: AFIT/ LSH (Thesis Feedback), Wright-Patterson AFB, Ohio 45433.

1. Did this research contribute to a current Air Force project?

   a. Yes b. No

2. Do you believe this research topic is significant enough that it would have been researched (or contracted) by your organization or another agency if AFIT had not researched it?

   a. Yes b. No

3. The benefits of AFIT research can often be expressed by the equivalent value that your agency received by virtue of AFIT performing the research. Can you estimate what this research would have cost if it had been accomplished under contract or if it had been done in-house in terms of man-power and/or dollars?

   a. Man-years _____ $ _____ (Contract).

   b. Man-years _____ $ _____ (In-house).

4. Often it is not possible to attach equivalent dollar values to research, although the results of the research may, in fact, be important. Whether or not you were able to establish an equivalent value for this research (3 above), what is your estimate of its significance?

   a. Highly Significant b. Significant c. Slightly Significant d. Of No Significance

5. Comments:

Name and Grade Position

Organization Location

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

NO POSTAGE
NECESSARY
IF MAILED
IN THE
UNITED STATES

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO. 73236 WASHINGTON D.C.

POSTAGE WILL BE PAID BY ADDRESSEE

AFIT/LSH (Thesis Feedback)
Wright-Patterson AFB OH 45433

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

14
AFIT

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: LSSR-83-80	2. GOVT ACCESSION NO.: AD-A093202
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): AN INVESTIGATION OF EXPERIMENT DESIGNS FOR APPLICATIONS IN BIOFEEDBACK-PERFORMANCE RESEARCH METHODOLOGIES,	5. TYPE OF REPORT & PERIOD COVERED: Master's Thesis
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Gilbert Fried, Captain, USAF	8. CONTRACT OR GRANT NUMBER(s):
9. PERFORMING ORGANIZATION NAME AND ADDRESS: School of Systems and Logistics, Air Force Institute of Technology, WPAFB OH	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 1771
11. CONTROLLING OFFICE NAME AND ADDRESS: Department of Communication and Humanities, AFIT/LSH, WPAFB OH 45433	12. REPORT DATE: September 1980
	13. NUMBER OF PAGES: 91
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office):	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

APPROVED FOR PUBLIC RELEASE AFR 190-17.

FREDRIC C. LYNCH, Major, USAF
Director of Public Affairs

23 SEP 1980

18. SUPPLEMENTARY NOTES

Air Force Institute of Technology (ATC)
Wright-Patterson AFB, OH 45433

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Biofeedback
Electromyogram
Experimental Validity
Experiment Design
Human Performance
Relaxation Training
Self-Regulation

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Thesis Chairman: Saul Young, Lt Col, USAF

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

For a number of years biofeedback techniques, to control physiological responses, have been gaining wide acceptance in the area of clinical research. A previous thesis experiment explored the possibility of using these techniques to enhance task performance; however, no significant results were obtained. The purpose of this thesis was to research the literature on experiment design techniques to gain a better understanding of experiment design complexity and, in so doing, find other avenues of approach which may lead to more significant results. The material covered included (1) five general experiment designs which are considered to be the basic "building block" techniques for numerous other hybrid designs; (2) experiment design with humans concentrating on the important concepts of internal and external validity; and (3) biofeedback experimental research efforts with an eye toward task performance and any relationships to the previous thesis effort. As a result of the investigation several shortcomings of the original experiment were pointed out and two recommendations were advanced for increasing the validity of any future attempts at biofeedback/task performance research.

LSSR 83-80

AN INVESTIGATION OF EXPERIMENT DESIGNS FOR APPLICATIONS IN BIOFEEDBACK-PERFORMANCE RESEARCH METHODOLOGIES

A Thesis

Presented to the Faculty of the School of Systems and Logistics

of the Air Force Institute of Technology

Air University

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Systems Management

By

Gilbert Fried, BS
Captain, USAF

September 1980

This thesis, written by

Captain Gilbert Fried

has been accepted by the undersigned on behalf of the faculty of the School of Systems and Logistics in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN SYSTEMS MANAGEMENT

DATE: 19 September 1980

Paul Young
COMMITTEE CHAIRMAN

Edward J. Dunn Jr.
READER

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation and thanks, first and foremost, to my wife Helene for accomplishing not only the typing of this thesis but also being able to "keep the home fires burning" through these long days and nights. Also, a special thanks goes to my twin sons, Mike and Doug, who endured many hours of uncharacteristic quiet while I pursued the completion of this effort.

I also thank my thesis advisor, Lieutenant Colonel Saul Young, for his guidance, strong encouragement, and friendship through this effort; and my reader, Lieutenant Colonel Edward J. Dunne, for his valuable insight and suggestions which aided in the successful completion of the thesis.

Finally, last but not least, a note of thanks goes to the School of Systems and Logistics librarian, Jim Helling and his staff, for their kind and courteous assistance in gathering research material.

TABLE OF CONTENTS

LIST OF TABLES

Chapter I

INTRODUCTION

The striving for increased performance is everpresent on the minds of top management whether it be in business and industry or in the United States Air Force. In the Air Force, mission performance is a primary goal and a significant portion of that goal may be reached by concentrating on improving human task performance. The human factors discipline has made great strides toward that end in the area of individual and machine interaction. But can individual performance also be improved by learning to control one's own physiological responses to task oriented situations?

Background

Eastern civilizations have, for many years, known techniques for voluntary control of physiological body functions; however, it was not until the social revolution of the '60s that any serious effort at scientific study was undertaken. As Hume wrote, in a 1976 review of biofeedback research (23:1), "Ten years ago anyone writing a review of research in this area would have had difficulty in filling a postcard with the relevant information." Along with the public's involvement in mind and consciousness raising, the scientific community's interest in the relatively young discipline of psychophysiology

emerged. This discipline is concerned with the complex relationship of mind and body. Psychologists held pro and con opinions concerning the possibility of modifying autonomic nervous system activity through operant conditioning methods (basically a reward system for behavior that induces a satisfactory response). To provide insights into the mechanism and scope of these mind-body interactions biofeedback research techniques were developed (8:v; 9:2).

Biofeedback is basically a process of "feeding back" information about an individual's physiology to the individual generating that information. The psychophysical technique involves monitoring physiologic activity about body functions such as heart rate, blood pressure, skin temperature, muscle tension, or brain waves. The information is "fed back" by visual and/or auditory means indicating changes in the activity. An individual then must develop a personal mental mechanism whereby voluntary control over the physiologic activity is exerted (9:3-4).

When biofeedback burst on the scene it was picked up and touted by the popular press as a natural panacea therapy for the cure of many forms of illness. This was done many times without adequately assessing the true value of the experimentation that was accomplished. One of the biggest problems in this area, partly due to the nature of the process is substantiating the claimed cures used in advertisements by biofeedback therapists, training schools, and equipment

suppliers. Fortunately, a significant amount of work has been done by competent researchers over the past 15 years. In 1969 the the Biofeedback Research Society was founded by a group of psychotechnologists interested in furthering research and applications in the field of self-regulation. With the group expanding, in 1976 the name was changed to the Biofeedback Society of America and now has 1100 members with eight state groups. The society publishes a quarterly journal, a quarterly newsletter, and proceedings of its annual meetings (10; 23:1; 39:v).

In a relatively short time period, with calmer heads prevailing, biofeedback has been recognized as having many worthwhile applications but also many limitations. No less an august body than the National Aeronautics and Space Administration (NASA) is presently investigating the use of biofeedback techniques for its space shuttle program. Astronauts are learning to control physiological attributes such as heart rate and blood pressure in an effort to reduce the effects of motion sickness (4:54).

As in the NASA investigation mentioned above, a look at several current publications (8; 9; 23; 34; 39) dealing with biofeedback indicates that a vast majority of the material deals with clinical research and applications. Areas of concern include hypertension, migraine headaches, tension headaches, heart rate disorders, and insomnia just to name a few. The investigation of biofeedback techniques as a task

performance enhancer has been relatively unexplored. The possibility of performance benefits gained from active use of biofeedback piqued the interest of Professor Young at the Air Force Institute of Technology (AFIT). This interest stemmed from the fact that Young had been involved with clinical applications of biofeedback for some time. That involvement resulted in a 1979 AFIT thesis by Kipperman (26) which attempted to show a relationship between the use of biofeedback techniques and performance.

The Experiment

Under an original concept of Young's, Kipperman's biofeedback research was concerned with studying the possibility of improving pilot performance. The experiment employed a single physical task of pitch tracking, with no induced motion, using a pilot control stick and a vertical motion display. The experimental environment was an enclosed Roll Axis Tracking Simulator supplied by the Aerospace Medical Research Laboratory at WPAFB. The biofeedback technique used was the electromyogram (EMG) which measures the amount of muscle electrical activity. This indication has been found to be an accurate representation of muscle tension level.

Biofeedback training and EMG feedback was conducted using electrodes placed on the subject's foreheads to measure the electrical activity of the frontalis muscles. The training consisted of one 30-45 minute session where relaxation and physiological control techniques were explained and practiced.

This was followed by a 5-10 minute reinforcement period before each tracking session which consisted of eight runs. Questionnaires were filled out to identify any differences in demographic variables that could affect the experimental outcome. The twenty male volunteers who participated were then assigned to three experimental groups. Seven received EMG feedback training and active EMG feedback during tracking, six received only EMG feedback training, and five received no training or active EMG feedback. The remaining two were considered expert trackers and after EMG feedback training received active EMG feedback on half of their tracking runs. Each subject was scored by computer, for tracking error, on at least 48 runs each having a duration of three minutes. The experimental analysis was accomplished using the data base of tracking error scores and corresponding EMG electrical activity values (26:2-3,5-6,16-17).

Analysis and Results

Multivariate data analysis techniques were employed using a standard statistical computer program package. The analyses used included linear regression, correlation, and analysis of variance (ANOVA) methods to test for significance of several entering variables on a particular criterion variable. A scatter diagram program was also used to provide visual plots of each subject's tracking error score versus run number. The data was grouped for analysis in the following three ways: (1) aggregated by individual, (2) aggregated by run number, and (3) individually (26:21).

The results of the analyses were not significant in terms of the major hypothesis proposed. The most conclusive result of the analyses performed was in an area not directly related to the experiment. The scatter diagrams of tracking error score versus run number showed that a logarithmic relationship in a learning model is valid. No significant differences between experimental groups were found when tracking error score was used as a single measure of performance. The experiment did not indicate that active EMG feedback or EMG feedback training was a significant predictor of performance enhancement. In fact the analyses showed that active biofeedback was associated with a lower rate of learning (26:39-41). The question then is, why were the results of the experiment generally inconclusive?

Statement of the Problem

In experimental research the design of the experiment has a major effect on the results achieved. This is an area of real concern and may well have been a cause for the lack of meaningful results. It has been postulated that a counteractive effect, of hearing a constant biofeedback signal, prevented full concentration on the tracking task and resulted in a degradation of performance (26:41). Another possibility for not achieving valid test results was the integration of the task learning experience at the same time the biofeedback data was collected.

An experimental methodology needs to be found that will indicate task performance improvement or degradation without the possible negative effects mentioned above. If a more valid experimental method can be found and utilized, the results obtained should clearly show any significance, or lack thereof, between subjects who have or have not received exposure to biofeedback techniques.

Objectives

Primary Objective

The purpose for this thesis is to investigate the available literature concerning experiment design techniques. The intent is to gain knowledge of other possible avenues of approach which may lead to increased validity of experimental results from research on task performance improvement through biofeedback techniques.

Secondary Objective

A second purpose for this study is an attempt to synthesize some of the work accomplished in the area of experiment design techniques.

Personal Objective

Through exploration in the area of experiment design I will, at a minimum, expand my awareness of the inherent complexity of the field and many of the methodologies employed. Armed with this knowledge, I can better evaluate the relative success alluded to in research reports.

Scope and Limitations

This thesis is intended to explore existing experiment design techniques with an orientation toward idiopathic studies. In terms of biofeedback research, idiopathic refers to a strategy that is peculiar to the individual in attempting to voluntarily control some physiological activity. In light of the primary objective, the research is limited, as much as possible, to areas involving task learning and task performance evaluation. The broad scope of the experiment design discipline and the time constraint inherent in this effort is a limiting factor in the treatment of experimert design and associated statistical methods. Also, this author is painfully aware of his limited knowledge in the area of statistics; therefore, except for the possible mentioning of techniques that may be applicable to particular experiment designs, any rigorous treatment of statistics underlying the experiment designs will be left for the statisticians to work through.

Methodology

This thesis is intended to provide a literature review of experiment design techniques with an eye toward the possible improvement in design validity of biofeedback task performance enhancement research. To accomplish this, first covered is some of the general experiment design literature to gain an overall insight in the field. Next, major emphasis is placed on investigating existing literature concerning specific problems in experiment design with humans. Following

that is a look into some of the current literature on biofeedback experimentation to discover what, if any, research has been accomplished in the area of task performance enhancement. Finally, using the experiment design information available, Kipperman's experiment is evaluated in terms of design methodology and validity factors with an eye toward promoting a more successful experiment methodology in biofeedback performance enhancement research.

Chapter I of this thesis has introduced the subject matter, provided background material, and the overall research approach. Chapter II comprises a discussion of the experiment design discipline as a whole.

Chapter II

GENERAL EXPERIMENT DESIGN

Chapter I, comprising the introduction to this thesis, discussed the biofeedback phenomenon and the task performance experiment using biofeedback technology. The lack of significance in the test results led to questions about the validity of the experimental methodology and thus to the primary objective of this effort. This chapter will cover experiment design techniques using a "broad brush" approach to gain overall knowledge of the subject. Later chapters will delve into narrower areas consisting of experiment design with humans and biofeedback experimentation. To introduce this chapter let us first look at the evolution of the science of experiment design.

Background

Federer and Balaam, in their discussion of the development of experiment design (15:40), first credit the mathematician L. Euler for his work in 1782 as "having a profound effect upon research on the construction and properties of experiment and treatment designs . . ." Also of some importance to modern experiment design was a paper written in 1832 that describes methods for statistical comparisons in dealing with the standardization of weights and measures. During the

same period research began in the area of agricultural methods which eventually led to a comprehensive theory of experiment design. As Finney (16:45) stated, "The first great stimulus to the development of the theory and practice of experimental design came from agricultural research." In fact, several of the standard terms used in experiment design today owe their origins to the work done in agricultural research (16:46).

Modern experiments in agriculture began in France, in 1834, when Boussigault undertook the first practical field experiments on his farm. In England, in 1841, John Bennet Lawes was instrumental in establishing the Rothamsted Experimental station on his farm. Joined by J. H. Gilbert in 1843, they spent almost the next six decades in carefully planning and executing field experiments. Many others toiled in the field of agricultural experimentation through the 19th and into the early 20th century; however, most of the field layout designs then in use resulted in conclusions that were difficult to validate. It is at this point, chronologically, that the name R. A. Fisher crops up in much of the background literature in experiment design (15:40; 16:45).

Fisher's realization of the problems in validating field testing results led him, from 1923 on, to study the underlying principles of scientific experimentation and eventually to develop new design techniques. His writings on the subject are of tremendous importance especially the book first published in 1935 entitled The Design of Experiments. No one can dispute the fact that other individuals have made

great advances in the science of experiment design beyond what Fisher has accomplished; however, his work is still recognized by many as the cornerstone in the field. The following quotes certainly attest to that fact. Federer and Balaam (15:40) remarked that "It might be stated that experiment design as known today, had its beginning with Sir Ronald A. Fisher," Kempthorne (24:vii) wrote that "His contributions to the Logic of the scientific method and of experimentation are no less outstanding, and his book The Design of Experiments will be a classic of statistical Literature.," and Finney (17:1) stated, "Fisher's book remains the classic that everyone having any concern with experimental design must read . . ." Thus, much of the experiment design discussion in this chapter, while drawn from many sources, relies heavily on the pioneering work accomplished by Fisher. But before that can be accomplished some terminology and definitions need to be explained so that the uninitiated may comprehend what is to follow.

Terminology and Definitions

It may be of interest to note that while pouring over the literature in the field no general consensus could be found as to whether "experiment design" or "experimental design" is the proper terminology. Federer and Balaam (15:8), no doubt concerned, saw fit to devote a paragraph to this problem. They, in essence, said that published literature is highly confused in use of the term experimental design. It

has been used in discussions of selection of sample size or treatments and conduct or analysis of experiments. Their opinion was that "experiment design" was a more appropriate term since it compares more favorably with other research terms such as treatment design and survey design. So, recognizing the astuteness of their argument, this author has opted to use "experiment design" throughout this thesis. In any event, regardless of which term one prefers, what does it actually mean? In a number of references (1:87; 15:1; 27:87) experiment design is defined rather succinctly as the method or approach whereby treatments are arranged or placed on experimental units. Finney (16:3) defines it more broadly as,

> (i) the set of treatments selected for comparison; (ii) the specification of the units (animals, field plots, samples of blood) to which the treatments are to be applied; (iii) the rules by which the treatments are to be allocated to experimental units; (iv) the specification of the measurements or other records to be made on each unit.

The terms mentioned above plus others basic to the science of experiment design will now be defined. Note that since a number of books researched on the subject describe and/or define terminology the definitions below have been constructed from several sources.

Alternative Hypothesis

A statement of value for a dependent variable that is phrased to contradict the null hypothesis is called an alternative hypothesis. The alternative hypothesis is usually the

one that a researcher hopes to affirm; however, it cannot be proven directly but is one that remains tenable if the null hypothesis is rejected (see Null Hypothesis), If the alternative hypothesis is accepted then it becomes possible to infer that the experimenter's original research hypothesis is true (27:23; 31:16).

Blocking

When available experimental units are not homogeneous in nature the inherent differences in characteristics tend to mask treatment effects. By grouping the units by their similarities into blocks under different treatments the sources of heterogeneity are effectively isolated. These units are then relatively homogeneous within each block and uncontrolled variations can then be measured by comparisons between them. A major purpose of blocking is to increase the power of an experiment to detect treatment effects which is accomplished by variance analysis. Blocking is one of the terms carried over from agricultural experimentation. In accomplishing field plot trials of soil treatments compact blocks of adjacent plots were used to control the effects of soil heterogeneity (1:86; 16:50; 25:31).

Confounding

When comparison tests can only be made for treatments in combination and not for a separate treatment in question, then the dependent variable effects are s id to be confounded.

Effects due to one treatment variable cannot be distinguished from effects of other treatment variables. Confounding is sometimes deliberate when an attempt is made to reduce the number of treatment level combinations assigned to experimental unit blocks. In many instances confounding also arises as an inadvertent imperfection of an experiment design. This ultimately has the effect of confusing the experimental results (25:58-59; 27:553).

Correlation

When a researcher is concerned with the strength of the relationship between variables, rather than the probability of observing a particular value, then the concept of statistical correlation is useful. The quantitative expression used to denote the extent of the relationship between two variables is the correlation coefficient. The values of the coefficient will vary between +1.00 and -1.00. Positively correlated variables increase or decrease in value in the same direction, with perfect positive correlation indicated by a value of +1.00. Negatively correlated variables increase or decrease inversely with respect to each other, with perfect negative correlation indicated by a value of -1.00. A value of zero indicates no correlation between variables. Usually, in inference testing, the null hypothesis is stated as zero correlation between variables. A rejection of the null hypothesis in favor of the alternative hypothesis will then indicate evidence for the existence of correlation (25:66; 31:18; 37:94-95).

Dependent Variable

A dependent variable is usually a measurable quantity that is observable but is not controlled by the experimenter. Rather, it is a chosen indicator that reflects the effects associated with manipulation of an independent variable or variables. The quantification is important so that a statistical analysis may be accomplished to make inferences about a research hypothesis (27:5; 31:10).

Experimental Unit

The smallest part of a sample population that is differentiated for the purpose of receiving some treatment is designated an experimental unit. In agricultural research a small area of land chosen for a treatment, with dimensions specified by an experimenter, is called a plot. The term has stuck and is used in other experimental disciplines to denote an ultimate experimental unit even though it may refer to something entirely different than an area of land (16:461; 27:555).

Hypothesis

Webster's dictionary (42:410) defines hypothesis as "a tentative assumption made in order to draw out and test its logical or empirical consequences . . . Hypothesis implies insufficiency of presently attainable evidence and therefore a tentative explanation." In experiment design literature this definition is typically treated in two distinct parts (see Research Hypothesis and Statistical Hypothesis).

Independent Variable (Treatment)

A variable that is under the control of the experimenter so that it can be manipulated to assume different values is termed an independent variable. The terms independent variable and treatment are used interchangeably. A broad definition proposed by Linton and Gallo (31:8) is that ". . . any variable, regardless of type, that is assumed to produce an effect on, or be related to, a behavior of interest." In practice, different values of an independent variable (treatment) are applied in order to confirm or deny the existence of differential effects on a dependent variable (25:296; 27:4-5).

Null Hypothesis

A researcher performing an experiment typically hopes to find differences in a dependent variable of experimental units that have either received or not received a particular treatment. The null hypothesis is a statement of value for that variable, phrased in such a way as to negate the possibility of a relationship between the treatment and dependent variable in question. The null hypothesis is the statement under test and is assumed to be true. If it is rejected under a test of significance then the alternative hypothesis can be accepted (see Alternative Hypothesis). To understand this reasoning it is important to bring up the notion of indirect proof. Runyon and Haber (37:166) describe it in terms of "the logic of statistical inference." The

null hypothesis, as an exact statement of value, can never be statistically proven. Also, the alternative hypothesis, as a mutually exclusive statement, is always proven indirectly through the rejection of the null hypothesis under an acceptable level of risk (see Test of Significance). As Fisher (18:16) points out,

> In relation to any experiment we may speak of this hypothesis as the "null hypothesis," and it should be noted that 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.

In other words, the conclusions made about any experiment are not absolute and remain open to further questioning. The one statement that can be made without any hesitation is that findings or hypotheses can never be asserted as true without any doubt (27:23; 31:15-16; 37:166-168).

Population

In statistical usage the term popuation refers to any finite or infinite collection of individuals or observations governed by an identifiable set of rules which determine membership. Population statistical parameters are seldom measured in experimental research due to the restriction of size; therefore, population parameters such as means or standard deviations are usually estimated by sample values drawn from the overall population and by applying appropriate statistical methods (27:2; 31:14).

Randomization

The idea of complete randomization is to allow each experimental unit an equal chance of receiving each possible treatment. If randomization can be accomplished any biases in the treatment effects resulting from uncontrolled error variations may be eliminated. The tests of significance on any dependent variable in question then can be considered valid. An important point to consider is that a so-called "haphazard" selection process can be fraught with unintended biases. One sure way to avoid this problem is to use a random number generator or table for assignment of treatments and experimental units. Also, it is not necessary to assure complete randomization if the variation introduced is known to be of negligible effect and will not vitiate the results. The relevance of this concept is evident by this statement of Lindquist's (30:11-12) that, "one of the most important and basic of all principles of experimental design is thus the principle of randomization" and Montgomery's (32:3) comment that, "Randomization is the cornerstone underlying the use of statistical methods in experimental design." As one of the significant contributions formulated by Fisher, randomization is considered one of the few truly modern characteristics of experiment design (1:86-87; 13:6-8; 16:32).

Replication

The meaning of replication is the collection of data from two or more observations with the experiment being performed under a set of identical experimental conditions. Of course this is an ideal situation and rarely happens in fact. Practically, a researcher will attempt to hold the variations in experimental conditions to a minimum. The basic experimental layout is held intact though changes in place and time may occur as the experiment is repeated. The intent of replication is to increase precision and gain a closer estimate of sampling error. With increasing replications comes an increase in experiment sensitivity and decrease in error of treatment comparisons (24:177; 25:249; 27:3; 29:72).

Research Hypothesis

The research hypothesis is a word statement that presumes a relationship between independent and dependent variables. The hypothesis may be derived from a theory, observations, or simply educated guesswork; however, as a word statement it cannot be tested directly but must be transformed into a mathematical value (see Statistical Hypothesis) that can be compared (27:3; 31:15).

Sample

A sample is a subset or part of a population made available by some process, usually deliberate selection, for the purpose of investigating particular properties of the

parent population. Random samples are drawn from populations of interest using the methods described for randomization (see Randomization). In practice though samples rarely ever meet the strict criteria for randomness. But they are treated as though they did if no systematic biases exist that could be expected to invalidate the results (25:254; 27:2; 31:15).

Statistical Hypothesis

The assumed research hypothesis forms the basis for the statistical hypothesis. The statistical hypothesis becomes a mathematical statement of a proposed value for one or more parameters of a population. Since the statistical hypothesis must be logically derived from the research hypothesis, its acceptance or rejection gives the researcher a basis on which to evaluate the truth or falsity of the initial research hypothesis. To reach this conclusion the experimenter must actually formulate two mutually exclusive and dichotomous statistical hypotheses (see both Null and Alternative Hypothesis) from the research hypothesis (21:12; 27:3,23; 31:15).

Test of Significance

The use of tests of significance allows a researcher to evaluate the probability that some observed sample value of a dependent variable would occur, given that the stated null hypothesis were true. Typically, a test statistic with a known probability distribution is employed to define the probability of that null hypothesis. One of the most powerful

and widely used statistical tests (pick up any book on statistics or experiment design and it will be discussed in detail) is the analysis of variance (ANOVA) technique. The set of procedures developed can be used for simple as well as complex experiments involving simultaneous comparisons of many variables. If the probability associated with the test statistic is sufficiently low then the null hypothesis can be rejected and the alternative hypothesis affirmed. If the opposite occurs then the null hypothesis cannot be rejected. Either way, the test of significance provides the researcher with statistical evidence for concluding that there are or are not true differences between dependent variables under different treatments (20:92; 27:4; 31:17,122).

It should be noted that the listing of terms was accomplished alphabetically for convenience only and therefore has no intended hierarchical effect. Also, it should be apparent that the listing is by no means exhaustive; however the terms defined here, as chosen from several sources, are those that were deemed important to a basic understanding of experiment design concepts. Other terms that may crop up in this and later chapters will be discussed as necessary when they occur. The next section includes a discussion of several experiment design methodologies.

Experiment Design Methodologies

In researching literature for this section this author was struck by the myriad of text books dealing with the subject of experiment design and related statistical concepts. This is evidenced by the bibliographical listing accumulated for this thesis which is but a small portion of the material available. One reason for the plethora of information seems to be that experiment design techniques cut across a wide diversity of physical and social sciences. Thus, researchers in each discipline have published works relating to experiment design with methodologies that are geared to studies in a particular science. With so many design approaches available the problem then was how to limit the scope to fit within the framework of this discussion of general experiment design.

After some thought, the solution that emerged was to attempt a "common thread" approach to reviewing the literature. The reasoning was that there would be some methodologies repeatedly covered that could then be considered as basic experiment designs and those would be the ones discussed here. Before getting into individual methodologies though, an important consideration for comment is the question of what constitutes a "good" experiment design?

Foremost to the accomplishment of a successful experiment is to plan in detail just what is actually required. This involves formulating a clear statement of the problem to be investigated. Once that step is done then the researcher

can determine the other interrelated activities that make up the complete experiment design. These include items such as setting up a statistical hypothesis to test, planning for the collection and analysis of data to test the hypothesis, defining the treatments to be applied, selecting the sample population to be investigated, and selecting the criterion (dependent) variable to indicate variation effects of the treatments (21:1; 27:1). As important considerations for the activities mentioned above Lindquist (30:6) summarizes and Kirk (27:22) essentially concurs with a number of essential characteristics of a good experiment design as follows:

> 1) It will insure that the observed treatment effects are <u>unbiased</u> estimates of the true effects.
>
> 2) It will permit a quantitative description of the precision of the observed treatment effects regarded as estimates of the true effects.
>
> 3) It will insure that the observed treatment effects will have whatever degree of precision is required by the broader purpose of the experiment.
>
> 4) It will make possible an objective test of a specific hypothesis concerning the true effects; that is, it will permit the computation of the relative frequency with which the observed discrepancy between observation and hypothesis would be exceeded if the hypothesis were true.
>
> 5) It will be <u>efficient</u>; that is, it will satisfy these requirements at the "minimum" cost, broadly conceived.

With that short excursion into what constitutes a "good" design concluded, the original focus of this section will now be continued. In searching through the literature for that "common thread," five types or approaches to experi-

ment design were discussed most often, not necessarily as simple basic designs, but as major constructs or building blocks for any of the most complicated designs. Both Kirk (27:11) and Lindquist (30:7) make essentially similar statements in that regard. The remainder of this chapter will therefore be devoted to descriptions of the following design methodologies; completely randomized, randomized block, Latin square, incomplete block, and factorial. Much of the presentation is based on information culled from references of Finney (16), Hicks (21), Kirk (27), and Lindquist (30); therefore, specific citations to these sources will not be made unless an important point is raised by an individual author.

Completely Randomized Design

Probably the simplest of all designs, in terms of both formulation and statistical analysis, is one in which the relevant treatment levels and experimental units (plots) can be randomly chosen for interaction. In the completely randomized design no restrictions are placed on randomization. What that means is that not only are plots randomly selected from the available sample population but also the treatments under investigation are randomly assigned to the plot groups. A typical layout for this design is shown in Table 1 and the symbolism used is defined as follows:

T_j = Treatment variable or level

X_{ij} = Random Treatment interacted with random plot

$\bar{X}._j$ = Mean of summed data under treatment T_j

TABLE 1

COMPLETELY RANDOMIZED DESIGN LAYOUT

T_1	T_2	.	.	.	T_j
X_{11}	X_{12}	.	.	.	X_{1j}
X_{21}	X_{22}	.	.	.	X_{2j}
.	.	.	.	.	.
.	.	.	.	.	.
X_{i1}	X_{i2}	.	.	.	X_{ij}
$\bar{X}._1$	$\bar{X}._2$	.	.	.	$\bar{X}._j$

The analysis is made by summing the plot data under each treatment variable and then performing hypothesis testing by comparing the means. The analysis usually consists of a one-way (single treatment variable) ANOVA with the null hypothesis stated as no difference between level effect means.

This design has several advantages. First, it is flexible in that any number of treatments and/or plots can be used. The number of plots can be varied from treatment to treatment, however the recommended procedure is to equally divide plots among the treatments; second, the analysis used is relatively easy even if the numbers of plots for all

treatments is not the same, or the uncontrolled sources of variation (experimental error) in the data differ from treatment to treatment; third, the method of analysis remains unchanged even when results from plots are missing and the relative loss of information due to unavailable data is less than with other more complex designs. The experimental error mentioned above leads to an important area of consideration for researchers and also to the principle criticism of the completely randomized design.

Experimenters always attempt to minimize error effects, both experimental and design, primarily because they contribute negatively to the accuracy of test results. Unfortunately, com›lete randomization fails to isolate or equalize individual plot variations. That has the possible effect of biasing the data in favor of one or another treatment effect. A way of eliminating this problem is to have large enough samples so that individual differences will tend to average out and the plots can be considered to be homogeneous. The solution sounds easy but in reality the researcher faces obstacles that for the most part are insurmountable. The most crucial ones are cost and time considerations for accomplishing the experiment, and the usually limited resources available from which to draw experimental units. So, ultimately upon judging the various considerations, the completely randomized design

is most appropriate where units under observation are homogeneous or where the accuracy is good enough to overcome the effects of variation error. Otherwise, it seems that an alternative design is required to obtain satisfactory experimental results.

Randomized Block Design

Both Cochran and Cox (13:107), and Finney (16:51) agree that the most frequently used of all experiment designs is the randomized block. This design is based on a principle of assigning individual plots to blocks so that the plots within each block are more homogeneous. The differences among the blocks are then considered as nuisance variables that are essentially isolated through the experiment design. The plots, in this case, are not randomly selected from the sample population but are grouped for homogeneity. Randomization is restricted within the blocks to assignment of treatment levels for individual plots. The basic design layout is shown in Table 2 with the symbolism essentially the same as for the completely randomized design. The "X_{ij}" symbol now indicates a block level plot interacting with a random treatment level and the symbol "$\bar{X}_{i}.$" is added which represents the mean of summed block effect data.

TABLE 2

RANDOMIZED BLOCK DESIGN LAYOUT

	T_1	T_2	.	.	.	T_j	
Block (1)	X_{11}	X_{12}	.	.	.	X_{1j}	$\bar{X}_1.$
Block (2)	X_{21}	X_{22}	.	.	.	X_{2j}	$\bar{X}_2.$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
Block (i)	X_{i1}	X_{i2}	.	.	.	X_{ij}	$\bar{X}_i.$
	$\bar{X}._1$	$\bar{X}._2$	.	.	.	$\bar{X}._j$	

Looking at the layout in Table 2, the differences among the column means ($\bar{X}._j$) represent treatment effects whereas the differences among the row means ($\bar{X}_i.$) represent block effects. Remember that the block effect indicates differences in the criterion or dependent variable due to pre-selected variations among the plots. The analysis is fairly straightforward and employs the two-way (treatment and block effects) ANOVA. With this design two test hypotheses can be investigated. Not only can possible treatment effects be looked at but also effects of plot variations can be studied. The main objective however, is usually to test for treatment differences.

The major advantage of this design over a completely randomized one is that it provides some control over the error effect. This is accomplished by avoiding total random assignment of plots to treatments. The testing of treatment effects becomes more valid due to the reduction of experimental error based on individual plot differences. In using a design that has the property of being able to reduce the error variance, the researcher can expect a more precise estimate of treatment effects. Thus, the randomized block design is more powerful than the completely randomized design if, in a particular experiment, block effects account for a significant portion of the total error variance. Here again though, the experimenter must weigh the added complexity of matching plots against the possible gain in precision. Quite possibly, the degree of precision is still not satisfactory and the search for other alternatives must go on.

Latin Square Design

When a researcher feels that two uncontrollable variables within plots represent major sources of experimental error then the possible use of a Latin square design is indicated. This design evolved from an ancient puzzle which involved finding the number of ways Latin letters could be arranged in a square table so that each letter appeared only once in each row and column. The blocking principle, as used

in the previous design, is extended in this design to achieve homogeneity among two so-called nuisance variables. The two variables are assigned in block levels to rows and columns of a square matrix whose size is determined by the number of treatments under investigation. Each treatment is randomly assigned to a cell within the square under two restrictions. Each treatment can appear in any row only once and in any column only once. Table 3 shows a Latin square layout for a case with three treatments or levels. The symbolism used is defined as follows:

A_i = First uncontrolled variable level

B_j = Second uncontrolled variable level

T_k = Treatment variable or level

X_{ijk} = Cell data point (random treatment T_k interacted with block A_i and B_j levels)

$\bar{X}_{i..}$ = Mean of summed data for A_i

$\bar{X}_{.j.}$ = Mean of summed data for B_j

$\bar{X}_{..k}$ = Mean of summed data under T_k

Note: To achieve a complete Latin square design there must be equal numbers of blocks and treatments.

TABLE 3

LATIN SQUARE DESIGN LAYOUT (3X3)

	Block (B_1)	Block (B_2)	Block (B_3)	
Block (A_1)	T_1 X_{111}	T_2 X_{122}	T_3 X_{133}	$\bar{X}_{1\cdot\cdot}$
Block (A_2)	T_2 X_{212}	T_3 X_{223}	T_1 X_{231}	$\bar{X}_{2\cdot\cdot}$
Block (A_3)	T_3 X_{313}	T_1 X_{321}	T_2 X_{332}	$\bar{X}_{3\cdot\cdot}$
	$X_{\cdot 1\cdot}$	$\bar{X}_{\cdot 2\cdot}$	$\bar{X}_{\cdot 3\cdot}$	

$$\bar{X}_{\cdot\cdot 1} = (X_{111} + X_{321} + X_{231})/3$$

$$\bar{X}_{\cdot\cdot 2} = (X_{212} + X_{122} + X_{332})/3$$

$$\bar{X}_{\cdot\cdot 3} = (X_{313} + X_{223} + X_{133})/3$$

From Table 3, the differences among the row and column means, ($\bar{X}_{i\cdot\cdot}$) and ($\bar{X}_{\cdot j\cdot}$) respectively, represent two different block effects. The differences among the means derived from the calculations in Table 3 ($\bar{X}_{\cdot\cdot k}$) represent treatment effects. The basic analysis used to test the two block hypotheses and one treatment hypothesis is simply an extension of the ANOVAs described previously; however the primary interest is still with evaluating the treatment effects. With proper planning

in the choice of variables to be blocked some of the effects of unwanted variation can be avoided. This does not totally eliminate the concept of randomization since, in the proper use of this design, treatments are randomly selected as stated earlier. This selection is accomplished by randomly choosing from all of the possible Latin squares of a particular size. A quick look at the limited data in Table 4, taken from Finney (16:58), indicates that as the square size increases the random possibilities become enormous.

TABLE 4

POSSIBILITIES FOR LATIN SQUARES

Size of Square	Number of Different Squares
2X2	2
3X3	12
4X4	576
5X5	161,280
6X6	812,851,200
7X7	61,479,419,904,000

The Latin square design has the overall effect of reducing experimental error even further than the two previous designs. A chief restriction on the use of the design is that as the number of treatment levels gets larger the multiplicity of experimental trials necessary to satisfy differential

cell requirements becomes impractical. Here again, the researcher must make a choice among alternative designs based on the experimental environment he faces.

Incomplete Block Design

A situation where an incomplete block design is particularly applicable is when the available plots for block levels turns out to be less than the number of treatments being investigated. Another case would be if the number of treatments was so large that to fill all the required blocks would mean a significant loss of precision due to the heterogeneity of the plots. The formulation of this design requires that each block have the same number of plots, each treatment occur the same number of times, and plots be assigned to treatments so that every possible pair of treatments occurs together within some block an equal number of times. If the symmetry described here is achieved then the design is considered balanced; otherwise the design is partially balanced. Of course, the closer the formulation can be made to total symmetry the higher the precision of means tests will be. A balanced incomplete block design layout for three treatments is shown in Table 5. The symbolism used is the same as that used for the randomized block design.

TABLE 5

INCOMPLETE BLOCK DESIGN LAYOUT (BALANCED)

	T_1	T_2	T_3	
Block (1)	X_{11}	-	X_{13}	$\bar{X}_{1\cdot}$
Block (2)	-	X_{22}	X_{23}	$\bar{X}_{2\cdot}$
Block (3)	X_{31}	X_{32}	-	$\bar{X}_{3\cdot}$
	$\bar{X}_{\cdot 1}$	$\bar{X}_{\cdot 2}$	$\bar{X}_{\cdot 3}$	

Differences among row and column means can be compared to determine block and treatment effects as in a randomized block design. But to reach the point of analysis with this design requires a high degree of computational dexterity. The statistical analysis is far more tedious than in the designs discussed previously. Initially what is required is the solution of a set of linear equations whose number is determined by the relationship (T + B - 1), where "T" is the number of treatments and "B" is the number of blocks. The computations then quickly become worse, so that is about all this author is willing to say about the statistical analysis at this point. Anyone wishing a more complete discussion is referred to Cochran and Cox (13:380-384,443-463) for the details.

A disadvantage to using this design is, in fact, the cumbersome computations that are required. Also, there are no formal procedures for accomplishing a symmetrical layout,

which makes it more difficult to construct the experiment. If the researcher, due to resource considerations, finds it necessary to use this approach then there are some benefits to be derived. This design does permit the evaluation of several treatments without the necessity of having complete blocks. Also, the design can be more efficient than a randomized block since the method of analysis gives not only intra-block information but inter-block information as well.

Factorial Design

This methodology is actually not a particular design but a number of techniques that combine other designs. A brief discussion is felt to be important though because factorial methods have gained widespread acceptance especially in the area of behavioral research. The method makes use of combinations of established building block designs to permit the evaluation of the effects of two or more different treatments simultaneously within a single experiment. Two of the most commonly used designs have been discussed earlier. They are the completely randomized and the randomized block methodologies. A researcher might consider a factorial experiment for two primary reasons. One, information concerning the average effects of each of several different treatments can be obtained from one experiment of moderate size. Two, this method allows the researcher to assess interaction effects among the treatment variables.

As in all things, there are advantages and disadvantages to selecting a factorial type of experiment. On the plus side, if the treatments are independent of each other, or their interactions can be considered insignificant, then comparison of the main effect of each treatment can be evaluated separately in the analysis. Also, one experiment can provide information on several treatments whereas several experiments with single treatment variables would be required. On the negative side, if the treatments are numerous then the size and complexity of a single factorial experiment can become most unwieldy. If interaction effects are strong then the analysis will be more complex because those effects must be accounted for in any ANOVA procedure. The interpretation of the results becomes more difficult because of treatment interactions.

Ultimately, it is the responsibility of the researcher to make an evaluation early on as to what is required to accomplish an experiment with satisfactory results. The approaches described here are only the very basic ones necessary for a cursory look at the subject. They can be combined and modified almost ad infinitum to suit any number of experimental environments. A key point to note is that an "expert" statistician can be of immeasurable help in determining what experiment design to follow since, in the end, it is the statistical analysis that will decide the

validity of the results. Also, randomization seems to play some part in all of the basic designs discussed, as a deterrent to biasing any data. As Finney (16:48) so eloquently stated, ". . . when in doubt, randomize."

This chapter has presented some background information on the experiment design discipline; some basic terminology has been described and defined; and the major building block designs have been discussed to give some feel for the basic approach to general experiment design. Armed with this basic knowledge, Chapter III takes a look at the unique problems arising in experiment design when human subjects are the experimental units under treatment.

Chapter III

EXPERIMENT DESIGN WITH HUMANS

In Chapter II the subject of experiment design was introduced. A number of historical points of fact were mentioned in relation to the development of the experiment design discipline. Following that discussion, several terms deemed basic to the topic were defined which led to descriptions of five types of experiment designs. The choice of particular designs was made based on a number of sources in the field referring to them as basic building blocks to the science of experiment design. With that as a basis, this chapter will narrow the field of view somewhat and discuss problems in experiment design validity as they typically occur when humans are the experimental units under treatment.

Background

More than in any other type of research, when human subjects are the units used for receiving experimental treatments the difficulties that arise in attempting to achieve valid test results increase in intensity. Probably the key discriminator between experimentation in the "hard" sciences (engineering, biology, agriculture, etc.) and those that primarily focus on human beings is heterogeneity. What is meant by that is, as experimental units, human subjects

exhibit an inherent multiplicity of diverse characteristics which tend to maximize between subject differences; whereas, in the "hard" sciences it is somewhat easier to achieve some semblance of homogeneity between units subjected to experimental treatments. This is so because it is much easier to maintain uniformity when dealing with inanimate materials, laboratory test speciments, chemical solutions, and the like.

What became apparent, through researching the literature for this section, is that there seemed to be vastly more written about experimentation in the social/behavioral sciences and education but not much in the area of human subjects interacting with a physical task environment. Parsons (33:1-2) however, in discussing the broader topic of man-machine system experimentation, proclaims that there are many characteristics in common with research in the behavioral sciences. Another point he makes is that the man-machine type of research is a hybrid of sorts and therefore is not claimed by the more traditional researchers as their own. This could possibly be a reason for the subject not being widely publicized. At any rate, using this tack of a strong relationship to behavioral studies, the following section concerning human subject design problems draws heavily on two sources from the behavioral area. The first is a survey of experiment designs for teaching research by Campbell and Stanley (11) which was initially published in 1963. The second is a text by Walizer and Wienir (41), recently published (1978), dealing

with research methods for the behavioral sciences. Both works contain fairly comprehensive treatments of experiment design and validity, with the latter borrowing much from the former. In point of fact, Walizer and Wienir (41:242) refer to the effort by Campbell and Stanley as ". . . the modern 'bible' on experimental design." Since most of the material for this discussion has been extracted from the two sources mentioned above, specific citations will not be made unless quotes or other references are used.

Experimental Validity

Webster (42:980) essentially defines validity as the quality or state of achieving a conclusion that is correctly derived from certain premises, or the state of being well grounded. In the context of experiment design, validity consists of two definable distinct concepts which are of primary concern to a researcher attempting to achieve satisfactory results. These two important concepts are internal and external validity. Internal validity refers to the criterion that, in fact, an experimental treatment is the causal factor for a specific set of experimental conditions. External validity refers to how extensively, beyond an experimental setting, can a treatment effect be generalized (11:5). The former term being the more critical of the two will be dealt with first.

Internal validity is considered an indispensible criteria and therefore an absolute necessity to obtain satisfactory experimental results. Without internal validity it is impossible to interpret any experiment to determine if a treatment made a difference within the sample population. To ideally achieve internal validity there are several extraneous variables (those not related to any treatment variables) that need to be controlled. The strong necessity for control is that, if not controlled, the effects of the extraneous variables can confound any treatment effects thereby confusing the test results. Campbell and Stanley (11:5) give short definitions of classes of extraneous variables as threatening to internal validity. Walizer and Wienir (41:243-247), using a few different titles, discuss similar classes of variables in somewhat greater detail. The substantive material below describes those extraneous variables as a synthesis from the two sources cited above.

History Effects

Events that occur, in addition to the experimental treatment, make up the history within a subject's experience. The events other than the treatment may produce unwanted history effects unless they are controlled by the researcher. Basically, there are two ways this can be done. The first is to totally isolate the experimental subjects so that the only event experienced is the treatment under study. This method

is impractical at best under most situations. The second is, by judiciously choosing a suitable experiment design, to negate the confounding effects of extraneous events.

Maturation Effects

There are many processes that may occur within experimental subjects that are simply the results of a temporal condition. Maturation involves the changes that take place over time other than those related to historical events. Some of these maturation effects that can contribute to confounding are growing older, getting hungrier, becoming fatigued (both physically and mentally), getting bored, and so on. It is obvious from these examples that the time period involved in the experiment is a determining factor for the onset of each condition. So it behooves the researcher to carefully consider a design choice in light of the length of the experiment to be run.

Testing Effects

In research methodologies that make use of pretests (measurements obtained prior to a treatment) and posttests (measurements obtained subsequent to a treatment) the post-test scores may be affected solely due to the fact that a pre-test was given. If a measurement involves a type of performance that tends to improve with repetition, then administering a pretest provides an opportunity for practice and a likelihood for higher scores, any treatment nonwithstanding. As Campbell

and Stanley (11:9) state, ". . . students taking the test for a second time . . . usually do better than those taking the test for the first time." An opposite effect, resulting in lower posttest scores, is also possible if a pretest causes fatigue, boredom, anxiety or other detrimental effects on a subject.

Reactivity Effects

Campbell and Stanley do not treat reactivity as a separate extraneous variable class but bring it up in relation to their discussion of testing effects. Walizer and Wienir however, do break out reactivity as a distinct source of confounding. It is being singled out here because reactivity seems to be a confounding variable of significant concern in the context of biofeedback experimentation. Reactivity effects surface any time human subjects knowingly participate in an experiment. The problem arises because subjects are aware of their participation and therefore may react differently than if they had received a treatment in a non-experimental setting. A key point, as Campbell and Stanley (11:9) put it, is that

> The reactive effect can be expected whenever the testing process is in itself a stimulus to change rather than a passive record of behavior. . . . In general, the more novel and motivating the test device, the more reactive one can expect it to be.

The way out of the dilemma is to use nonreactive measures if possible. However, that likelihood is relatively remote. Only if the experimental setting commonly occurs in the subject's normal environment can the absence of reactivity effects be assured.

Instrumentation Effects

When using measuring instruments in an experiment, the calibrations may change which will introduce an undetected effect on a dependent variable measurement. This instrumentation phenomenon can also occur when human observers are the "measuring instruments." In fact, the variability of a human observer is probably greater and more likely to occur than if mechanical instrumentation is used. Recognizing that this effect will happen, it can be controlled by seeing that any instrumentation or decay occurs equally among experimental groups.

Statistical Regression Effects

In a pretest-posttest methodology where a researcher is interested in studying subjects with extreme initial scores the problem of obtaining unreliable posttest measures exists. This is so because, as a statistical phenomenon, when multiple measurements are taken scores tend to regress toward some mean value. For example, if a group were selected for treatment based on their extremely low measurements and then retested, any observed increase may not have been due to the treatment. Some of the initial low scores may have been chance events so, on average, some increase would be expected to occur simply due to the regression effect. Here again, the proper choice of a design can control the possible confounding of the extraneous variable.

Selection Effects

The preferred method for controlling differences in experimental subject's characteristics is through selection by randomization to treatment and non-treatment groups. If randomization is not done then a selection effect occurs which obscures any possible conclusion that could be made from observed differences between treatment and non-treatment groups. The selection effect is based on the subject's characteristics, in each experimental group, being the cause for the observed difference. In essence, each group is different with or without any treatment and attempting to equalize groups by matching characteristics is not considered sufficient. The problem with that approach is that it is always possible for a researcher to neglect an important matching characteristic or conversely, to match subjects based on some insignificant ones.

Experimenter Bias Effects

Walizer and Wienir (41:247) single out experimenter bias as an extraneous variable with a strong influence on confounding. The effect occurs when an experimenter, either knowingly or by chance, somehow influences a subject's response in an experiment. Bias can also result when an experimenter has knowledge of the hypothesis being tested and/or which subjects are members of treatment or non-treatment groups. There are a number of techniques commonly employed in an attempt

to reduce the effects of experimenter bias. These include such methods as: (1) having the persons responsible for giving instructions or taking measurements ignorant as to the hypothesis under study, (2) insuring that the same persons in (1) above also are not aware of which subjects are in treatment or non-treatment groups, and (3) using automated devices to record data whenever possible. The first two methods are usually expressed as the experimenter being "blind" to the hypothesis and/or subject manipulation. Anyone wishing to explore this subject further is referred to Rosenthal's (35) expansive text titled Experimenter Effects in Behavioral Research. Also, the other side of the coin (subject bias) is treated in depth in Rosenthal and Rosenow's (36) book The Volunteer Subject.

Subject Mortality Effects

When there is a differential loss of subjects in a comparison group experiment, or a loss of subjects in a single group pretest-posttest design, mortality confounding effects are possible. There are a multitude of reasons for such losses including illness, moving, or just plain not showing up. Rosenthal and Rosenow (36:23-25) discuss the issue of subjects not appearing for appointments and such individuals are referred to as "pseudovolunteers," a term coined by other researchers (apparently, it seems, researchers feel better when phenomena are given neat little names). When mortality effects occur, differences in measurement means after a

treatment may be due to the loss of subjects whose individual differences affected pre-treatment measurements. There is no ideal way to protect against this extraneous variable regardless of the type of experiment design being used. The best a researcher can do, when faced with a differential loss of subjects, is to see that the perceived important individual characteristics of the remaining subjects are balanced.

In addition to the specific variables relevant to internal validity, as discussed above, there are combinations of effects. These are somewhat more obscure and involve interactions between selection effects and others mentioned above such as maturation, testing and history. Should all of the sources of confounding be ideally controlled then an experiment is considered internally valid; however, the experimenter must also be concerned with external validity as well. To what population, outside of the subjects measured, can a treatment effect be generalized? This is a more nebulous concept to consider as compared to internal validity. Generalization or representativeness of an experiment involves being able to infer that what happens to a sample population can be extended to some larger population. As in internal validity, there are several factors that can jeopardize external validity and these will be discussed here.

Testing and Treatment
Interaction Effects

An experiment design that involves a pretest measure may present a threat to external validity due to a testing and treatment interaction. The effect occurs because, in the "real" world, measures taken from subjects to establish a baseline tend to affect their response to a treatment variable. As a result posttest measures are not representative of any possible treatment effect on the population under investigation. Controls for this effect may be achieved by either eliminating the pretest measure entirely or by choosing an experiment design that will account for the anticipated confounding variable.

Selection and Treatment
Interaction Effects

The selection of subjects is an important consideration for both internal and external validity. To achieve internal validity the preferred method is to randomly assign subjects to treatment groups; however, if the researcher is interested in a specific population group this presents a problem for external validity. The subject groups selected in this manner eliminates any guarantee that they are representative of any particular group of interest. To eliminate this effect randomization of subject selection must occur from all individuals in a population of interest. This

is very difficult to achieve due to considerations such as geographical location, subject population availability, and time and cost factors.

Experimental Arrangement Effects

One of the most difficult dilemmas to try and resolve is whether or not experiments in a controlled or natural setting produce the best results. In many cases this may be a moot point due to the nature of the treatment under investigation requiring complex instrumentation and controls. A natural setting tends to increase external validity but extraneous variables are more difficult to control thereby threatening internal validity. In any event, internal validity, being the more critical of the two, should not be sacrificed to satisfy external validity. If at all possible, conducting like experiments in both settings would allow a determination of whether or not results are obtained on the basis of arrangement effects alone.

Multiple-Treatment Interference Effects

Typically, in experiments where a single group of subjects are presented with multiple treatments, the results are representative only to an overall population in terms of the same sequence of multiple treatments. This happens because, in general, the effects of one treatment do not

disappear and therefore have some residual effect on the results of another treatment. The resulting interference effect clouds the evaluation of each treatment with respect to its generalizability to other than the sample population.

It is evident from the discussions above that in order to obtain valid test results the sources of confounding and the factors affecting generalization must be considered very carefully. With that in mind, the next section will discuss some experiment design methodologies in behavioral research and how they relate to experimental validity.

Human Subject Experiment Designs

In Chapter II the point was made that the number of different design approaches was almost limitless. Here too, in an area where the researcher has to concern himself with not only a multitude of experimental situations but also the idiosyncrasies of human behavior, the design possibilities are limited only by the researcher's inventiveness. Campbell and Stanley chose a limited number of designs and most expertly discoursed on the nature of each design; and how it either satisfied or did not satisfy both internal and external validity factors. To reiterate all of their work here would be a pointless waste of time since they said it better and with more knowledge than this author ever could; however it is important that the reader get a feel for how validity factors do interact with experiment designs.

Campbell and Stanley divided the designs they discussed into three groups as follows:

1. Pre-Experimental designs that suffer more, in relation to other designs, by not being able to satisfy many of the validity factors.

2. "True" experimental designs that are the ones most recommended in much of the experiment design literature.

3. Quasi-Experimental designs that exhibit a lack of complete experimental control over such things as randomization of human subjects and the scheduling of treatments.

One experiment design of each category will be presented out of Campbell and Stanley's (11) treatise as sufficient material to get at the crux of the issue in this section. Should the reader wish to investigate this topic in greater detail the original text is the recommended source.

Before proceeding with the experiment design discussion, there are some symbols that require definition and the design layout convention as presented in Table 6 needs to be addressed. The symbols used are defined as follows:

X = Represents exposure of a subject or group of subjects to a treatment variable whose effects are to be measured

O_i = Represents an observation or measurement process to gather data for evaluation of treatment effects

R = Represents the randomization process of assigning subjects to different treatment groups as the method of achieving pre-treatment equality

Each design layout is to be read from left to right indicating the temporal order of events. Any symbols that appear vertically oriented indicate that those events occur simultaneously.

One-Group Pretest-Posttest Design

A pre-experimental design that is still used in educational research, despite its faults, is the one-group pretest-posttest. The design layout is shown in Table 6. The design's simplicity and the possibility that no other alternatives exist are the most likely reasons for its use; however, several rival hypotheses, concerning extraneous confounding variables, become viable explanations of an O_1-O_2 difference. Validation of a treatment effect hypothesis which states that "X" is the cause for the difference is definitely a problem here, and how this design's results are confounded by extraneous variables will be taken up next.

Table 6

EXPERIMENT DESIGN LAYOUTS

One-Group Pretest-Posttest		O_1	X	O_2
Pretest-Posttest Control Group	R	O_1	X	O_2
	R	O_3		O_4
Nonequivalent Control Group		O_1	X	O_2
		O_3		O_4

Events that occur, other than "X," between the O_1-O_2 measurements rival the test hypothesis as producing the change. A key point is that this history effect will decrease the validity of the results as the time lapse between the two measurements grows longer. Experimental isolation is a cure but in the human behavioral arena it is practically impossible to achieve. Maturation effects are also a problem, independent of external events. In this single-group design no distinction can be made between the internal temporal processes of subjects and the treatment variable when trying to determine a causal relationship for a measurement difference. Testing effects confound the results as explained in the previous section. The pretest itself then, most likely affects the posttest measurements and becomes a rival hypothesis for any

measurement change. Along with testing effects reactivity to the pretest measure can also influence the posttest scores adding to the confusion.

Instrumentation effects are another cause for the lack of control when using the single-group pretest-posttest design. Changes in the measuring device may account for a detected O_1-O_2 difference. The effect is more likely when human observers are the means for measurement; however, when automated equipment is used the effect is still present in calibration errors. Regression effects, as discussed earlier, are a hazard in this design when subjects are chosen for a treatment specifically because of the extremity of their pretest measurements. A change in measurement means can be mistakenly attributed to "X" when in fact the extreme pretest scores have simply regressed naturally toward the mean. It is apparent, from this discussion, that this experiment design leaves a lot to be desired. In fact, this design does not conform to any of the major building block designs discussed in Chapter II. Each of the building block designs generally require more than one treatment group for comparative analysis whereas this design deals with a single group and a single treatment mode.

Pretest-Posttest Control Group Design

In an effort to control the confounding effects of extraneous variables affecting internal validity, a control

group can be added and equivalence of groups achieved through randomization, resulting in the pretest-posttest control group design. The layout of this"true" experiment design is shown in Table 6. This methodology is one of the most frequently used and is called the ". . . classical experimental design . . ." by Walizer and Wienir (41:231) who go on to say that it ". . . is considered the grand master of research designs because in a relatively simple way the researcher can deal with many of the problems of demonstrating causation. . . ." This design conforms to the completely randomized design, as discussed in Chapter II, where it is possible to randomly select subjects for different treatments. History, maturation and testing effects are all controlled through being manifested equally in both the treatment and non-treatment groups. Care must be taken to avoid intrasession history effects caused by simultaneous experimenter differences. To achieve a balanced representation for this and other biases randomization of experimental occasions should be followed if possible. A point to be made here is that the design, using an experimental group receiving "X" versus a control group receiving no "X," is an oversimplification. The control group may in fact be experiencing other levels of "X" or a different set of activities altogether which in reality adds some ambiguity to any evaluation of the effect of "X."

Regression and selection effects are essentially controlled through the randomization process assuring group equality. In the case of regression, if both treatment and control groups are randomly selected from the same extreme population then they should equally regress toward the mean on posttests. Thus, the O_1-O_2 difference can still be related to the application of "X" alone. Selection effects are ruled out by the same randomization of the subject groups; however, a proviso here is that true random samples based on statistical probability must be achieved. The larger the number of random assignments from an overall population the greater the assurance of group equality.

Instrumentation effects are easily controlled if intrasession history effects can be controlled; however, the problem is more difficult if observers are the instruments of measurement. The types of controls needed then, are those mentioned in the last section under the heading of experimenter bias effects. Mortality effects are controlled by the nature of the data collected within the design framework. When subjects in either experimental or control groups drop out, valid inferences about a O_1-O_2 difference can still be accomplished. This is done by retaining data for analysis from all subjects that have completed both the pretest and posttest. The apparent effect of "X" may be reduced but it does eliminate differential sampling bias. The pretest-posttest control

group design does a good job of controlling extraneous variables to give it internal validity; however, there are some concerns when the discussion turns to external validity.

The previous section pointed out that achieving external validity is somewhat more difficult than achieving internal validity. The hazards discussed there play a definite role in affecting the external validity of the pretest-posttest control group design. The pretest makes a testing and treatment interaction a strong possibility unless the nature of the testing is familiar and is seen frequently by the experimental subjects. Where highly unusual test procedures are employed it might be wise to choose an experiment design that does not contain a pretest. A selection and treatment interaction is also a strong possibility unless the tested subjects are able to be randomly selected from the entire population under study; otherwise, the results may only be valid for the specific groups measured. By far, the most pervasive of all threats to generalization is the experimental arrangement effect. Because of the very nature of experimentation, most attempts to control significant variables leads to an artificial setting. Regardless of the type of design, reactive effects are unavoidable unless some semblance of a natural setting can be achieved.

The "classic"pretest-posttest control group design does exhibit some difficulties with satisfying generalizability; however, as stated earlier, being internally valid is of

sufficient importance that its relevance to experimentation is still maintained. As good as the design is though, a key point is the requirement for randomization of subjects. On many occasions this is not possible and "untrue" designs must be used to accomplish the necessary research. One such experiment design and its relationship to validity will be explored next.

Nonequivalent Control Group Design

Quasi-experimental designs are used where designs that have a better "track record" in achieving validity are not feasible. It should be remembered, as discussed in Chapter II, that there are always risks in the testing of hypothesis. While these risks are somewhat greater in the quasi-experimental versus "true" designs they are still worthy of consideration. The researcher simply must be more wary of threats to validity when interpreting the results. The nonequivalent control group design is of the quasi-experimental variety and is widely used in educational research. As indicated in Table 6, it is similar to the "true" pretest-posttest control group design but the experimental subjects are not necessarily assigned randomly to the comparison groups. With a design of this type, the usual method used to achieve comparability of subject groups is to match them based on

individual attributes. A methodology similar to this one was described in Chapter II as the randomized block design. The dashed line represents comparison groups that are not equated by randomization.

History, maturation, testing, and instrumentation effects are controlled in a similar fashion as in the "true" design discussed previously. The degree of control is somewhat determined by how similar the treatment and non-treatment group are, based mainly on the similarity of pretest measures. The distinguishing internal validity factor of this design versus the "true" design is the threat of selection interactions with other extraneous variables. The lack of randomization in the selection of comparison groups may cause an O_1-O_2 difference that could be mistaken as the effect of "X." Regression effects are also a problem with this design because of inevitable differences in pretest group means, again due to the lack of randomization.

Though a few problems to internal validity were noted above, this design does fare a little better in controlling for effects that threaten external validity. The problems encountered in this area, for the most part, are similar to those presented in the discussion of the pretest-posttest control group design. Reactive arrangement effects however, are not as likely to threaten external validity in this design as in most "true" designs. The reasoning is that

where random selection for treatments is not used, naturally occurring groups of subjects are less likely to be aware of experimental manipulations.

The material presented in this chapter included a short background section containing some general comments concerning experiment design with human subjects; a discussion of factors relevant to internal and external design validity; and a description of three experiment designs and their interactions with threats to validity. Admittedly the brevity of the treatment does not do the subject justice; however, it is felt that the intent, to acquaint the reader with the complexity and difficulty associated with experiment design validity using human subjects, has been served. With Chapter II and this chapter as a foundation, Chapter IV looks at some of the literature involving biofeedback experimentation in the area of task performance enhancement.

Chapter IV

BIOFEEDBACK EXPERIMENTATION AND TASK PERFORMANCE

In the introductory chapter the question posed was, can task performance be improved using biofeedback control? The results of Kipperman's research was inconclusive and therein lay the purpose for this study. Chapter II dealt with areas of concern in the field of general experiment design. Chapter III narrowed the discussion to concerns where human subjects are the focus of experimental treatments. Both chapters have hopefully laid a foundation for an appreciation of the inherent complexities involved in devising a valid approach to biofeedback experimentation. This chapter looks at available literature involving biofeedback experimentation and any relationships those studies might have with respect to the Kipperman experiments, and suggests possible alternatives.

Background

In Chapter I, the point was made that most of the published material in the area of biofeedback concerns clinical research and applications. Though a couple of research reports

dealing with task performance were unearthed in the search, not a single reference was found relating biofeedback to performance in a non-clinical sense. On the clinical side, a number of books have been published geared to the public at large. These include Brown's (7: 9) popular works New Mind, New Body and Stress and the Art of Biofeedback; and Blanchard and Epstein's (6) A Biofeedback Primer. In the professional arena periodicals, such as Biofeedback: Research and Therapy and Biofeedback and Self Regulation, have long been established to report on biofeedback research and applications. The stature of biofeedback as an applied science has also made inroads in the medical profession. This is evidenced by the inclusion of works by Gaarder (19) and Basmajian (2) in the Wright State University School of Medicine library. Both books relate procedures and practices for the clinical use of biofeedback techniques. Because of this proliferation of work in the clinical area, most of the information presented in this chapter, of necessity, draws heavily from those efforts.

Here again, as discussed in Chapter II, the problem of what material to present had to be broached. The reader, using the bibliography as presented, could certainly gain more detailed information by going directly to the sources. Thus, the approach taken was to selectively discuss material that would whet the reader's appetite and at the same time relate to the Kipperman experiment. In that light, the next

section looks into biofeedback research as a scientific tool by including some general experiment design approaches.

General Biofeedback Experimentation

Biofeedback, as a valid scientific mechanism for promoting voluntary physiological changes, has been a long time in coming. Historically, there has always been a reluctance on the part of some to accept methodologies that smack of control especially when instrumentation equipment is involved. As examples, vocal resistance was strong when B. F. Skinner's pioneering work on behavior modification emerged, and some misunderstood the meaning of control as forwarded by Norbert Wiener when he proposed the now heralded cybernetic theory of systems (38:2). As a result, researchers shied away from investigating control of internal processes using instrumentation. Also, as mentioned in Chapter I, there were many who believed that the autonomic nervous system would not be responsive to voluntary operant conditioning.

Over the last two decades the opposition to biofeedback, as a science to be reckoned with and not a "fad" cure for all ills, has worn away as greater numbers of respected researchers entered the field. While there are still questions as to the efficacy of some of the methodologies employed, there can be no question that, at least in the clinical area, there are definite benefits to be derived from the use of biofeedback. With this in mind, in 1974 Blanchard and

Young (5) wrote an excellent review of published reports on clinical applications of biofeedback training employing various methods of feedback. Their study revealed a wide diversity of experimental procedures in use which they categorized into several groups. They professed, what was likewise stated earlier in this thesis, that the validity and reliability of conclusions concerning treatment effects hinges on the ability of an experiment design to control for extraneous variables. Short summaries of Blanchard and Young's (5:4-6) groupings are presented below to give the reader the gist of the types of experiment designs most often used in biofeedback research. They are described in the order of their ability to promote increasingly valid results.

Anecdotal Case Report

The weakest of all experimental methods reported is the anecdotal case report. No systematic collection of data is used but some description of a subject's clinical symptoms before and after the administration of a treatment is recorded. Also, some information is kept concerning the treatment itself and how it is applied to the subject. The value of this design lies not in its ability to produce valid conclusions, since it fails to control for most of the confounding variables discussed in Chapter III; but in its simplicity which allows for a minimum of effort and may suggest positive directions to take in accomplishing more rigorous research.

Systematic Case Study

A design somewhat better than the previous one described is the systematic case study. Here, data are collected through systematic measurement from a pre-treatment or baseline condition and several experimental or treatment trials. This design also suffers from a lack of control for several extraneous variables; however, with certain conditions it can yield acceptable results. These include the establishment of lengthy baseline data and experiencing a change in the symptom of interest that is coincident with the application of the treatment. While not a "true" experiment design, increased validity can be obtained through replication with several subjects. If similar treatment effects occur at approximately the same point in the trials then additional evidence for the efficacy of the treatment can be inferred.

Controlled Single Subject Experiment

A stronger design which uses an analysis of behavior is the controlled single subject experiment. Data are systematically collected across a minimum of three conditions. These consist of baseline, treatment, and return to baseline measurements. The return to baseline or reversal effect, as it is referred to, is considered the critical step. If a symptom change occurs with the application of a treatment and then returns to a baseline level when the treatment is removed, then evidence exists that the treatment is a causal variable for the symptom change. To increase the evidence

further, the treatment can be applied again to test for a second similar symptom change. Again, without many of the controls mentioned in Chapter III, any evidence gathered must be treated with caution.

Single Group Outcome Study

A common design approach, due to its simplicity, is the single group outcome study. Measurements of a target symptom are obtained from a similar group of subjects both before and after a treatment. Problems associated with a design of this type were discussed previously; therefore, the reader is referred to Chapter III where the information can be found under the One Group Pretest-Posttest Design heading.

Controlled Group Outcome Study

The most effective design represented in the biofeedback research literature is the controlled group outcome study. This design has also been discussed in Chapter III in two forms as the Pretest-Posttest Control Group ("true") Design and the Nonequivalent Control Group (quasi-experimental) Design. Although Blanchard and Young use Campbell and Stanley as a reference, their discussion of this design does not make the important distinction of randomization. They simply refer to the design as requiring a minimum of an experimental (treatment group) and a control (non-treatment) group of comparable subjects measured at the same time.

A modification of the controlled group design, and probably the least used due to its complexity and difficulty of operability, is the three group controlled experiment design. In addition to treatment and non-treatment groups an attention-placebo control group is added. Placebo is defined in Stroebel and Glueck (40:20) as ". . . any medication (treatment) used to alleviate symptoms, not by reasons of specific pharmacologic action, but solely by reinforcing the patient's favorable expectations from treatment." In a clinical environment, where therapy is administered to effect a change in a target symptom, placebo or expectancy effects are strong confounding variables which must be controlled. This has been a well known fact in drug or psychological therapy research and certainly extends to the area of bio-feedback experimentation.

Blanchard and Young's review of biofeedback research studies included, in the main, those that used EMG, heart rate, blood pressure, or electroencephalogram (EEG) as the feedback methodology. In summary they concluded, based on the soundness of the experimental procedures employed to yield meaningful clinical results, that EMG feedback methods yielded the most valid results (40:34). Based on this study, at least empirically, the choice of EMG feedback in the Kipperman experiment was a good one. With the preceding discussion as a base, the next section focuses on a few selected biofeedback studies to provide a look into some methodologies that have been used.

Biofeedback and Performance

A synopsis of two experiments and one long-term study concerned with different aspects of biofeedback research are presented below. The first experiment deals with the effects of relaxation training on mental performance under stress. The second experiment deals with the theraputic effects of EEG and EMG feedback for symptom reduction during detoxification from a methadone habit. The long-term study is concerned with several biofeedback modalities and their possible use in enhancement of human performance. The three bodies of work were selected to point out ramifications with respect to experimental validity and to serve as an aid in judging the efficacy of the Kipperman experiment.

Relaxation Training and Performance Stress

Chaney and Andreason (12:677-678) wished to determine the effects of a program of Jacobsonian progressive relaxation techniques (using specific voice instructions) on performance in a memorization test while under induced stress. Both galvanic skin response (GSR), a measure of skin electrical conductivity, and EMG feedback were used as dependent measurement variables of interest throughout the experiment. Forty-eight female student subjects were divided into matched triplet groups based on the following data: (1) college scholastic aptitude test scores; (2) Taylor anxiety test scores; (3) EMG masseter (lower jaw) muscle baseline tension levels;

(4) pretest memorization task tension levels; and (5) memorization task scores. Five one-way ANOVAs, performed on the matching variables, indicated no significant differences. The inference was made therefore, that each triplet was reasonably homogeneous prior to the experimental treatments. The triplets were then randomly selected for three possible treatments as follows:

1. A control group in which the subjects received no treatment for reducing tension but participated in a program of body mechanics.

2. An attention-placebo control group in which the subjects received a placebo pill daily that supposedly reduced tension.

3. The experimental group in which the relaxation techniques were taught with EMG visual and auditory feedback to monitor the progress of relaxation control.

All groups received six weeks of their respective treatments at which time a posttest was administered along with an induced stressor. A threat of academic grade point failure was made if no improvement was shown in the posttest versus pretest memorization scores. Quantitative measures were obtained on the same five variables used on the pretest, and ANOVA techniques were basically used in the analysis. The results indicated a significant difference in EMG levels, on the posttest, between the no treatment control group and

the experimental group; however, no significant difference was detected between the no treatment control group and the placebo group. The inference made was that the relaxation training group was able to control neuromuscular tension better than the other control groups when exposed to a stressful situation. Analysis of the memorization test scores indicated similar results although they were not quite as statistically convincing.

On the surface this experiment appears to be well thought out. The methodology used is a randomized block with a modification of the nonequivalent control group design discussed in Chapter III; however the threats to validity were not discussed by Chaney and Andreason as part of their results. The lack of randomization leads to selection interaction effects and regression effects due to the use of matching. Both could have a causal effect on differential posttest measures instead of the experimental treatment. Kipperman (26:6) assumed randomization when in fact, with the limited number of subjects available, it was not possible. Matching on subject attributes was required to obtain some comparability of treatment groups. It would seem then, that the same threats to internal validity were probably present. Experimenter and subject biases are a problem in the experiment discussed here and in the Kipperman study. Differential treatments used for both experiments were readily apparent to

both the researcher and the subject; therefore, biases for the successes or failure of one treatment or another are inevitable and become a detriment to validity.

As mentioned earlier, a placebo effect can be a strong confounding variable when human cognitive processes are at work. This factor was not addressed in Kipperman's work except in a passing concluding comment. Kipperman (26:41-42) remarked that ". . . people are different . . . One cannot expect different individuals to react the same way to the same situation . . . that fact must be incorporated in any analysis of experimental results." Even before analysis, individual differences need to be looked at in terms of the experiment design itself, controlling for not only placebo effects but bias and selection effects as well. Chaney and Andreason attempted to control for this effect but the methodology seems faulty. It seems, to this author, that to effectively control for placebo effects the researcher must be able to simulate the actual experimental treatment being tested. The difficulty in accomplishing this, when relaxation training and biofeedback is involved, is explored in the discussion of the next experiment.

A Double-Blind Methodology

Cohen, et al. (14:603-608) investigated the possible theraputic effects of EMG biofeeback for reducing the symptoms associated with detoxification from a methadone habit. Without going into the details, the results of their initial research, where both the experimenter and subject were aware of the treatments given (non-blind research), were less than conclusive. Many of the subjects who did not "learn" the EMG technique of tension reduction, based on collected EMG data, nevertheless were successful in reducing their detoxification symptoms. An apparent success bias was operating indicating a need for an attention-placebo control for this effect. That is easier said than done! Two major obstacles stand in the way of accomplishing the task. First, how do you keep the experimenter blind as to which subjects are receiving true or simulated feedback when the experimenter must be intimately involved with the biofeedback training procedure? Second, how do you present simulated feedback to the subjects, that will be essentially indistinguishable from true feedback? How this double-blind experiment design was ingeniously accomplished by Cohen, et al. in their second phase research is discussed next.

An additional group of subjects received simulated or actual biofeedback under the following test conditions:

1. Pre-constructed tapes of EMG feedback were obtained from subjects who, based on data in phase one, successfully accomplished tension reduction.

2. A tape recorder was connected to a computer and punch-card reader which was, in turn, hooked to the biofeedback apparatus. The system was set up so the computer could discriminate punch-cards which would allow either activation of the pre-recorded feedback tape, or actual real time feedback to be heard by the subject.

3. The experimenter's feedback display always indicated actual real time subject responses, regardless of which mode the subject was receiving.

4. The punch-cards, that controlled the treatment selection, were randomly distributed to the subjects who fed them into the card reader at the start of each trial.

With the above experimental controls instituted, the major difference between the control and experimental groups was the administration of real time or simulated feedback. Debriefings of both experimenters and subjects revealed that either group had no better than a chance probability of determining which treatment was being used. Analysis indicated that active biofeedback subjects in both the non-blind and double-blind phases were equally effective in achieving "learned" EMG tension control; while, little or no learning was achieved by the placebo subjects in phase two. The

success rate for reduction of detoxification symptoms however, was approximately equal for both the real time and simulated feedback groups. The results indicate a very definite suggestion that a placebo effect was a contributing factor. Cohen, et al. did caution others though, that the experiment was done under strict conditions with a very defined specific population and the research should therefore not be construed as valid for other clinical purposes. Of primary significance though, is the demonstrated feasibility of a double-blind technique within the severe constraints of biofeedback research. There is a potential here, given the necessary resources, for use in other situations such as the type of task performance experiment attempted by Kipperman.

A Five-Year Research Program

A computer search of the Defense Documentation Center files, relating to biofeedback research, turned up one study (3) concerned specifically with biofeedback as an aid to human effectiveness. The same study was reported on by Lawrence and Johnson (28) under the title Biofeedback and Performance. The five-year research effort, accomplished from 1970 through 1975, involved the efforts of 16 researchers. They performed a number of experiments involving brain activity, cardiovascular activity, muscle relaxation, and vasomotor activity. Their goal was to evaluate whether learning self-regulation of various physiological variables could enhance performance or

at least reduce performance loss under stressful situations. The general charter given the researchers was: (1) to train subjects to control various physiological variables, as mentioned above, using biofeedback techniques; (2) to look at correlations between the physiological variables and performance tasks; (3) to induce some form of stressor to observe its effect on the tasks, and (4) to determine if, through the learned self-regulation process, performance could somehow be enhanced. The experimental hypothesis tested in each case was that individuals who learn to control their internal physiology will perform better and have more control of behavior under stress (3:3-4).

The experiments and detailed procedures accomplished by the 16 researchers are far too voluminous to reconstruct here. It will suffice to say that many of the experiments involve fairly rigorous controlled conditions. The interest here is the overall results that bear on Kipperman's findings. Should the reader be interested in the details, a look at the original study is recommended. Several emphatic conclusions were arrived at upon the completion of the five-year study. In a number of laboratory studies, learning to regulate some internal physiological process through biofeedback training was accomplished in most cases; however, it was difficult, if not impossible, to alter the process to an endstate that was contrary to an individual's best interest in regard to their

to their own physiology. This points to the fact that individual differences play a large part in any form of biofeedback training or therapy, which makes generalization of experimental results exceedingly difficult (3:20).

In general, the results of several of the induced-stress experiments point to a lack of ability of subjects to maintain control over previously learned physiological responses when faced with a difficult task. The work accomplished in brain wave, heart rate, vasomotor, and EMG feedback indicated a lack of any significant relationship to performance enhancement. At the end of these studies, as of December 1975, the general opinion of the researchers involved was that further efforts to discover a biofeedback-performance link would prove fruitless (3:21-22). Kipperman (26:3-4) was aware of the negative conclusions of the above studies when he embarked on his experimental effort; however, the feeling was that additional efforts in this area were still warranted. His intent was to include more emphasis on learning and EMG measurement to search for a relationship between tension levels and performance. The experiment design however, did not allow for any pretest measure of performance to establish a baseline for the experimental and control groups. The temporal environment perhaps forced a compression of the biofeedback learning effort to coincide with the task learning effort, thus introducing treatment interaction effects. These

items, coupled with the experiment design difficulties mentioned earlier (placebo and selection effects; experimenter/subject biases), probably contributed to the lack of significant results in the Kipperman effort.

This chapter has presented a few background remarks relating to biofeedback experimentation; a discussion of the types of experiment design most often employed in biofeedback research; and a few selected studies that helped to point out some of the apparent difficulties in the Kipperman experiment. To be sure, there are numerous accounts in the available literature, especially in the clinical area, relating to biofeedback research; however, owing to the scope of this effort the choice had to be limited. Chapter V, as the finale of this thesis effort, briefly summarizes the material that has been presented to this point and ends with some conclusions and recommendations.

Chapter V

SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

Chapters II, III, and IV have hopefully led the reader to recognize some of the important relationships between experiment design in general, experiment design with humans, and biofeedback experimentation. In the process, these efforts were also intended to lead to evaluating the lack of significant results of the biofeedback experiment accomplished by Kipperman. This chapter presents a brief summary of the preceding work, any conclusions drawn from it, and closes with a few recommendations.

Summary

Chapter I introduced the topic with a discussion of the developing interest in biofeedback research over the past two decades. A discussion of Kipperman's experiment, which attempted to relate biofeedback and performance, followed pointing out the lack of significant results. The remainder of the chapter was taken up with defining the objectives of this thesis and the methodology to be followed. In the main, the intent was to learn about experiment design and in the process point out apparent problem areas with the Kipperman research effort with the hope of suggesting possible improvements.

Chapter II discussed some background information, in general terms, concerning the development of experiment design as a science. Specifically, the work of R. A. Fisher was singled out as the recognized pioneer in the field. Basic terminology was defined and discussed as a requirement for understanding the design descriptions to follow. Subsequently, five experiment designs were selected and discussed for their general treatment in the literature as the major building blocks in the field. These designs included completely randomized, randomized block, Latin square, incomplete block, and factorial. The layouts and usefulness of each design were presented to give the reader some idea of what is involved in the various approaches.

Chapter III then narrowed the field down to experiment design research involving humans as subjects. A background section related how the complexities of experiment design increase when heterogeneous human subjects are used versus the generally inanimate objects in the "hard" sciences. Most of the literature relating to experiment design with humans was found in the social/behavioral and educational sciences. Experimental validity, both internal and external, as major concepts for the success of any experiment were discussed in some detail. Several effects, as threats to either internal or external validity, were reviewed using mainly the work of Campbell and Stanley as the source. Finally, three experiment designs for research with humans

were discussed as representative of distinct stages of design effort. These stages are defined by Campbell and Stanley as pre-experimental, "true" experimental, and quasi-experimental designs.

Chapter IV, using the material presented in Chapters II and III as a foundation, approached the main issue fostering this effort. The area of concern, at the outset, was biofeedback and task performance and Kipperman's efforts to find a relationship between the two. A few background remarks were presented pointing out the relative lack of task performance literature versus clinical literature in the area of biofeedback research. A discussion of several experiment design techniques most often employed in biofeedback research followed. These techniques ranged from the simplest anecdotal case report to the most complex controlled group outcome study. The last section included a synopsis of three studies selected to aid in pointing out some apparent shortcomings in Kipperman's experiment.

Finally, of course, this chapter brings to a culmination the efforts presented up to this point. The following conclusions and recommendations sections hopefully indicate that the objectives initially expounded upon in the introductory chapter have indeed been somewhat successfully achieved.

Conclusions

As a result of the investigation undertaken for this thesis several apparent shortcomings in the Kipperman experiment were noted and may well have contributed to the lack of significant results. These factors are enumerated below.

1. Selection effects are probable due to the lack of randomization of subjects into treatment groups. Matching on attributes may be satisfactory if the matching produces closely comparable groups. The small number of subjects (twenty) makes comparability increasingly difficult. Also there was apparently no pre-experiment analysis of attributes or a pre-test to establish baseline data for matching.

2. Experimenter bias effects are probable since Kipperman was the sole experimenter throughout the research period. He was aware of all of the treatments presented, to which subjects each was presented, and recorded all of the associated data.

3. Subject bias effects are probable since each individual was aware of what type of treatment was being administered to them. Tracking task scores were also available after each trial plus EMG levels if they wished.

4. Though beneficial effects of biofeedback did not develop in the subjects, the placebo effect in research of this kind should be controlled. As mentioned in Chapter IV, the placebo effect is usually strong in any therapy oriented

research with human subjects, and biofeedback is primarily an internal therapy of self-regulation.

5. A problem alluded to in the introduction, the constant biofeedback signal affecting concentration on the task, may very well have contributed to a degradation of performance causing part of the erroneous results. Most of the experimental literature details the process of biofeedback training as a long term effort which is usually discontinued when "learned." The technique then became a purely internal one.

6. External validity problems should be fairly obvious from the discussion in Chapter III. Reactive arrangement effects are inescapable with the unique equipment involved and selection and treatment interactions are present by drawing subjects from a very limited sub-population (AFIT students) to represent the world of pilots.

The conclusions noted above, for the most part, do not reflect on Kipperman's competence as a researcher per se but on the actual methodology used. As mentioned early on, in Chapter II, there are several constraints that experimenters must deal with that affect the experiment design approach. Many of them, including items such as time available, cost factors, and availability of experimental subjects are factors that weigh heavily on AFIT student research. On a personal note, this author has gained much as a result of researching the literature for this thesis--certainly an overwhelming

AD-A093 202 AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL--ETC F/G 6/16
AN INVESTIGATION OF EXPERIMENT DESIGNS FOR APPLICATIONS IN BIOF--ETC(U)
SEP 80 G FRIED
UNCLASSIFIED AFIT/LSSR-83-80 NL

awareness for the complexity and difficulty of what is involved in achieving a valid experiment design has been driven home. With that thought in mind, the last section addresses a few provisional recommendations.

Recommendations

This author respectfully suggests the recommendations below should further attempts be taken to research relationships between biofeedback and task performance. It should also be realized that these recommendations come not from many years of experience but solely from gleaning the research material for this thesis.

1. A thorough training program, to learn how to voluntarily control a physiological response, should be undertaken prior to subjects performing a task. Experimental subjects should be pretested for this control ability without the necessity for an active biofeedback signal. Researchers (9:15; 39:150), at least in the clinical arena, point out that in order for voluntary control of internal responses to be effective in the "real world" individuals must be able to elicit this control without requiring constant biofeedback signals.

2. The double-blind methodology, described in Chapter IV, along with some form of pretest-posttest placebo control group design seem to hold the most promise for satisficing (achieving objectives within reason) experimental validity. Recognizing the complexity and scope of effort

required to accomplish a design of this type, it is this author's opinion that it is beyond the capacity of an AFIT student, given the limited time and resources available. Perhaps a member of the faculty as the researcher, with AFIT students as "blind" experimenters, and drawing subjects from a more stable population would be more suitable.

Probably, other experiment designs could be used with varying success; some better, some worse. The researcher, in the final analysis, must weigh all the factors to determine a "best" design approach and then correctly use statistics as an aid in "proving" or disproving any hypothesis. The researcher must protect against faulty designs and hypotheses; otherwise, all sorts of invalid proofs may materialize. As a final note, witness this old story related by Hooke (22:94)

> . . . about a flea trainer who claimed that fleas hear with their legs. As proof, he taught some fleas to jump at his shout of "Jump!" After amputating the fleas' legs and observing that they no longer responded to his shouts, he rested his case.

SELECTED BIBLIOGRAPHY

A. REFERENCES CITED

1. Anderson, Virgil L., and Robert A. McLean. Design of Experiments: A Realistic Approach. New York: Marcel Dekker, Inc., 1974.

2. Basmajian, John V., ed. Biofeedback: Principles and Practices for Clinicians. Baltimore MD: Williams and Wilkins Company, 1979.

3. Benshoff, Robert L., Project Leader. Self-Regulation as an Aid to Human Effectiveness and Biocybernetics Technology and Behavior. Annual report for the Defense Advanced Research Projects Agency. San Diego: San Diego State University Foundation, 1976. AD A021105.

4. "Biofeedback Helps Astronauts Fight Off Motion Sickness," Dayton Daily News, 103, No. 288 (June 1980), p.54.

5. Blanchard, Edward B., and Larry D. Young. "Clinical Applications of Biofeedback Training: A Review of Evidence." in Leo V. DiCara, et al., eds., Biofeedback and Self-Control 1974. Chicago: Aldine Publishing Company, 1975.

6. _____ and Leonard H. Epstein. A Biofeedback Primer. Reading MA: Addison-Wesley Publishing Company, 1978.

7. Brown, Barbara B. New Mind, New Body. New York: Harper and Row, 1974.

8. _____, ed. The Biofeedback Syllabus; A Handbook for the Psychophysiologic Study of Biofeedback. Springfield IL: Charles C. Thomas, 1975.

9. _____. Stress and the Art of Biofeedback. New York: Harper and Row, 1977.

10. Butler, Francine, Executive Director, Biofeedback Society of America, University of Colorado Medical Center, Denver CO. Telephone discussion. 23 June 1980.

11. Campbell, Donald T., and Julian C. Stanley. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally College Publishing Co., 1966.

12. Chaney, D. S., and L. Andreason. "Relaxation and Neuromuscular Control and Changes in Mental Performance Under Induced Tension." Perceptual and Motor Skills, 34 (February-June 1972), pp. 677-678.

13. Cochran, William G., and Gertrude M. Cox. Experimental Designs. 2d ed. New York: John Wiley and Sons, Inc., 1957.

14. Cohen, H. D., et al. "A Double-Blind Methodology for Biofeedback Research." Psychophysiology, 14, No. 6 (November 1977), pp. 603-608.

15. Federer, Walter T., and Leslie N. Balaam. Bibliography on Experiment and Treatment Design Pre-1968. New York: Hafner Publishing Company Inc., 1973.

16. Finney, D. J. Experimental Design and its Statistical Basis. Chicago: University of Chicago Press, 1955.

17. _____. An Introduction to The Theory of Experimental Design. Chicago: University of Chicago Press, 1960.

18. Fisher, Ronald A. The Design of Experiments. 6th ed. New York: Hafner Publishing Co., 1951.

19. Gaarder, Kenneth R., and Penelope S. Montgomery. Clinical Biofeedback: A Procedural Manual. Baltimore MD: Williams and Wilkins Company, 1977.

20. Hays, William L., and Robert L. Winkler. Statistics: Probability, Inference, and Decision. Vol. II. New York: Holt, Rinehart and Winston, Inc., 1970.

21. Hicks, Charles R. Fundamental Concepts in the Design of Experiments. New York: Holt, Rinehart, and Winston, 1964.

22. Hooke, Robert. Introduction to Scientific Inference. San Francisco: Holden-Day, Inc., 1963.

23. Hume, Wilfrid I. Biofeedback: Research and Therapy. Montreal, Canada: Eden Press, 1976.

24. Kempthorne, Oscar. The Design and Analysis of Experiments. New York: John Wiley and Sons, Inc., 1952.

25. Kendall, Maurice G., and William R. Buckland. A Dictionary of Statistical Terms. 2d ed. New York: Hafner Publishing Company, 1960.

26. Kipperman, Captain Mark C., USAF. "Use of Electromyogram Information to Improve Human Operator Performance." Unpublished master's thesis. GSM/SM/79D-18, AFIT/EN, Wright-Patterson AFB OH, December 1979.

27. Kirk, Roger E. Experimental Design: Procedures for the Behavioral Sciences. Belmont CA: Brooks/Cole Publishing Company, 1968.

28. Lawrence, G. H., and L. C. Johnson. "Biofeedback and Performance," in Gary E. Schwartz and Jackson Beatty, eds., Biofeedback: Theory and Research. New York: Academic Press, 1977.

29. Lee, Wayne. Experimental Design and Analysis. San Francisco: W. H. Freeman and Company, 1975.

30. Lindquist, E. F. Design and Analysis of Experiments in Psychology and Education. Boston: Houghton Mifflin Company, 1953.

31. Linton, Marigold, and Philip S. Gallo, Jr. The Practical Statistician: Simplified Handbook of Statistics. Monterey CA: Brooks/Cole Publishing Company, 1975.

32. Montgomery, Douglas C. Design and Analysis of Experiments. New York: John Wiley and Sons, Inc., 1976.

33. Parsons, Henry McIlvaine. Man-Machine System Experiments. Baltimore MD: The Johns Hopkins Press, 1972.

34. Proceedings of the Biofeedback Society of America; March 3-7, 1978; Albuquerque, New Mexico. Denver CO: Biofeedback Society of America, 1978.

35. Rosenthal, Robert. Experimenter Effects in Behavioral Research. New York: Irvington Publishers, Inc., 1976.

36. _____ and Ralph L. Rosenow. The Volunteer Subject. New York: John Wiley and Sons, Inc., 1975.

37. Runyon, Richard P., and Audrey Haber. Fundamentals of Behavioral Statistics. 2d ed. Reading MA: Addison-Wesley Publishing Co., 1971.

38. Schwartz, Gary E., and Jackson Beatty, eds. Biofeedback: Theory and Research. New York: Academic Press, 1977.

39. Stern, Robert M., and William J. Ray. Biofeedback; How to Control Your Body, Improve Your Health, and Increase Your Effectiveness. Homewood IL: Dow Jones-Irwin, 1977.

40. Stroebel, Charles F., and Bernard C. Glueck. "Biofeedback Treatment in Medicine and Psychiatry: An Ultimate Placebo?" in Lee Birk, ed., Biofeedback: Behavioral Medicine. New York: Grune and Stratton, 1973.

41. Walizer, Michael H., and Paul L. Wienir. Research Methods and Analysis: Searching for Relationships. New York: Harper and Row, 1978.

42. Webster's Seventh New Collegiate Dictionary. Springfield MA: G. and C. Merriam Company, Publishers, 1972.

B. RELATED SOURCES

Akey, Denise, and Nancy Yakes, eds. Encyclopedia of Associations; Vol. 1 National Organizations of the U. S. 14th ed. Detroit MI: Gale Research Company, 1980.

Barlow, D. H., et al. "Single Case Designs and Clinical Biofeedback Experimentation." Biofeedback and Self-Regulation, 2 (1977), pp. 221-240.

Beatty, Jackson, and Heiner Legewie, eds. Biofeedback and Behavior. New York: Plenum Press, 1976.

Birk, Lee, ed. Biofeedback: Behavioral Medicine. New York: Grune and Stratton, 1973.

Butler, Francine. Biofeedback: A Survey of the Literature. New York: IFI/Plenum Data Company, 1978.

Collier, Raymond O., Jr., and Thomas J. Hummel, eds. Experimental Design and Interpretation. Berkeley CA: McCutchan Publishing Corporation, 1977.

Cox, D. R. Planning of Experiments. New York: John Wiley and Sons, Inc., 1958.

Edwards, A. L. Experimental Design in Psychological Research. 4th ed. New York: Holt, Rinehart and Winston, 1972.

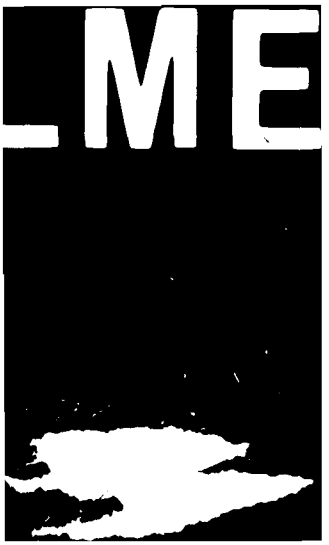
Myers, Jerome L. Fundamentals of Experimental Design. 2d ed. Boston: Allyn and Bacon, Inc., 1972.

Peng, K. C. The Design and Analysis of Scientific Experiments. Reading MA: Addison-Wesley Publishing Company, 1967.

Plutchik, Robert. Foundations of Experimental Research. 2d ed. New York: Harper and Row, Publishers, 1974.

BIOGRAPHICAL SKETCH OF THE AUTHOR

Captain Gilbert Fried enlisted in the Air Force in August 1968 and performed duties in the weather observer career field. Through the Airman Education and Commissioning Program he graduated from the University of New Hampshire, in January 1973, with a Bachelor of Science degree in Mechanical Engineering. After completing Officer Training School, he served as a test project manager with the 6511th Parachute Test Squadron, National Parachute Test Range, El Centro, California, through March 1977. His next assignment was with the 6514th Remotely Piloted Vehicle/Cruise Missile Recovery Test Squadron, Hill AFB, Utah, through May 1979. Following the Air Force Institute of Technology graduate degree program, Captain Fried will be assigned to the 6585th Test Group, Holloman AFB, New Mexico, as a Flight Test Project Officer.


LME