INTRODUCTION

Marvin L. Hanson, Ph.D.

I have strong conviction that clinicians, provided with some basic concepts, have greater potential for producing meaningful clinical research than do sophisticated researchers without clinical experience. There can be no doubt in our minds regarding the present need for sound research to test the effectiveness of myofunctional therapy. Traditionally, much of this research has come from colleges and universities. Funding for the investigations usually has been based on the relevance of the research to the training program. Since publication of the "Joint Committee Statement", which discouraged training of oral myologists until more research was done, such training programs apparently have been reluctant to invest time or money in this questionable area.

The IJOM is attempting to draw up a set of standards and policies that would make it possible for clinicians to pool their data into a nationwide research project concerning the efficacy of therapy for tongue thrust. The difficult part of this kind of research is the inclusion of a control group. Two problems immediately present themselves: (1) The choice and application of appropriate statistical procedures; and (2) the ethical questions of withholding treatment from a group of patients who need it.

Monte Smith has written four articles which provide insight into both these problems. Dr. Smith's series of articles will help us with both problems. I think you will find them very readable, understandable and helpful.

Experimental Evaluation of Social Programs

OVERVIEW STATEMENT

Almost a decade ago a social psychologist and methodologist by the name of Donald T. Campbell eloquently sounded a call to arms:

> The United States and other modern nations should be ready for an experimental approach to social reform, an approach in which we try out new programs designed to cure specific social problems, in which we learn whether or not these programs are effective, and in which we retain, imitate, modify, or discard them on the basis of apparent effectiveness on the multiple imperfect criteria available. (Campbell, 1975, p. 71; Originally published in the *American Psychologist*, April, 1969.)

Campbell urged the utilization of the experimental method to rigorously and objectively assess the effectiveness of various strategies designed to "cure" specific social problems. Regardless of the nature of the social problem, Campbell expressed a conviction that the social action program could be evaluated experimentally, provided that sufficient planning was provided prior to program inception. In other words, it made no difference whether the problem was juvenile delinquency or tooth decay, dyspedagogia or myodynamic dysfunction; given adequate technical resources, cooperation, and the opportunity to plan prospectively, the success of programs designed to ameliorate these problems could be evaluated experimentally.

The experimental evaluation of social programs, however, has proven to be more challenging than early proponents anticipated. Standard methodological problems encountered in laboratory research often are exacerbated in field settings, and accompanied by a host of problems indigenous to service delivery systems. Political realities, administrator resistance to random assignment, resentment and suspicion of the program evaluator; all these can be serious impediments to the rigorous evaluation of service delivery programs.

For those interested in assessing the effectiveness of service delivery systems, the experimental approach is not a panacea. An experimental assessment of a social program is difficult to implement, requires considerable planning and extensive cooperation among staff members, and often requires outside technical expertise. But the payoff is impressive. Relative to other evaluative methods, the experimental method will provide unambiguous answers. A carefully designed and implemented experiment permits the relatively unambiguous determination of program effectiveness.

Notwithstanding the numerous obstacles, rigorous experimental evaluations of social action programs are possible. The problems are not insuperable. The majority of impediments to successful experimental evaluations lie within the realm of what Ferman (1969) described as the dimension of social interaction. If staff members are willing to cooperate, experimental assessments can be accomplished.

This is the first of four articles dealing with experimental assessment of service delivery programs. The remainder of this article is designed to familiarize the reader with the important concepts of internal and external validity, and to demonstrate how "true" experimental designs differ from pre-experimental designs. Article number 2 in the series will review experimental, pre-experimental, and quasi-experimental designs most commonly employed in assessing social programs. The third article will discuss issues related to Dr. Hanson's category: "ethical questions of withholding, treatment from a group of patients who need it". The last article in the series will attempt to synthesize previous material and give practical suggestions for designing an evaluation strategy.

Reading and understanding these articles requires no formal training in research methodology or statistical analysis. Conversely, these articles alone will not make you a competent researcher. They will give you an introduction to the problems and procedures entailed in rigorously evaluating a service delivery program.

INTRODUCTION

Many people are unnecessarily intimidated by statistics and experimental design. I find this expecially true of those who in the course of their work find it necessary to interact with a research pseudosophisticate fond of bantering about such terms as analysis of variance, factor analysis, multiple regression, quasi-experimental design and other comparably esoteric appellations. Some of you may be interested, moreover, to know that each year I encounter advanced graduate students in psychology, often with several statistics courses to their credit, who are similarly intimidated.

Why the intimidation? Admittedly, there is a bewildering array of statistical tests, the mathematical bases of which are complex. Similarly, there is an incredible hodgepodge of experimental designs. Consider the consternation of one of my graduate students recently when she discovered that a given experimental design may be designated (equally correctly) by all these terms:

1) Linquist Type I Design
2) 3X4 Factorial Analysis of Variance (ANOVA) with Repeated Measures on the Second Factor
3) Split-plot Factorial 3.4 ANOVA
4) 3X4 Mixed Design
5) 3X4 with One Between and One Within Factor

Notwithstanding the admitted complexity and proliferation of colorful terminology, there is no compelling reason why anyone should be intimidated by statistics and experimental design. This article, therefore, serves as a nontechnical and nonintimidating introduction to how statistics and experimental design (predominantly the latter) may be used to assist in distinguishing between statistics and experimental design, proceed to a consideration of the relative merits of two approaches to assessing the efficacy of an hypothetical innovative health care program, and conclude with a tour of recent publications of these and related topics for the reader interested in more information.

Along the way I was flippant a time or two, and occasionally I simplify considerably (space is a consideration), but if you read and comprehend the remainder of this article you will come away with a rudimentary appreciation of experimental design application in the evaluation of social programs. Besides, along the way you will meet such interesting professionals as Don Design, master methodologist, and Dr. Cracker from State Teacher's University, classical statistician non pareil.

Distinctions and Definitions

Experimental design and statistics (Scylla and Charybdis of the research scientist's armamentarium) are not one and the same, but quite separate and distinct. In research practice the two are closely interrelated, but in terms of learning about the two, it is my contention that they can be considered separately, at least initially.

An experimental or research design is a systematic strategy for assigning experimental units to treatment conditions, manipulating variables, and making observations or taking measurements of other variables. In other words, an experimental design is a plan (or template) which can be superimposed over our research activities to insure that the little numbers (sometimes called data) which we collect can be interpreted unambiguously.

What are statistics? Remember the little numbers which our experimental design permitted us to collect? Well, Research Avenue, the road to unambiguous interpretation of those little numbers, traverses an intersection with Statistics Boulevard. At this intersection we ask questions of our friends, the little numbers. The interrogation may be prolonged, but there are formal procedures to ensure that the numbers are treated fairly. They are inspected closely to determine if they are fully qualified to answer questions. (Alas, some researchers have obtained answers from unqualified numbers. These are unscrupulous researchers and a pestilence to science.) If the numbers are fully qualified in all respects, then the numbers are induced to answer honestly through the application of predetermined operations. (We manipulate the numbers in prescribed ways.) These manipulations or operations are often referred to as statistical tests. Statistics then, are tools which we apply to our numbers (collected with the aid and comfort of our ally, experimental design) to enable them to answer our questions. A statistical test, then, is a tool, permitting a gaggle of little numbers to answer our questions.

There are many, many statistical tests (with a modicum of ingenuity we might manipulate our set of little numbers in an infinite variety of ways), and to study and master the application of all these statistics requires several years of intensive training. Even the most sophisticated of the statistical tests, however, do not provide us with unambiguous information unless our numbers have been collected under the overarching protection of a good experimental design. In this very important sense experimental design is not only distinct from statistics, but it is also more fundamental. Experimental design is more fundamental in the following sense. If you have collected a set of data (little numbers) within the proper constraints provided by a good experimental design, then you can always locate (either by trial and error or by contracting with an "expert") an appropriate statistical manipulation that will permit you to ask questions of your little numbers and derive therefrom unambiguous answers. However, if your data were collected without proper and careful recognition of the appropriate experimental design considerations (that is, if your study was sloppy), then notwithstanding the application of mountains of the most sophisticated and technically convoluted statistical tests, you can never tease unambiguous and clearly interpretable statements from your little numbers.

Of course, the best of both worlds would be to possess expertise in both statistics and experimental design, such that both design and analysis considerations can be made together, with due consideration for the proper fit. Such a course of training requires years, but it is possible, in a relatively brief time, to grasp the basic principles of sound experimental design. And that is my objective for the remainder of this article, to provide a simplified introduction to the fundamentals of experimental design, with the recognition that statistical considerations are a separate (but interrelated) issue.

Acknowledgement and Notation

The sections that follow are influenced greatly by the work of D. T. Campbell. The code of graphic presentation developed by Campbell and Stanley (1966) will be maintained herein, to enable the interested and enterprising reader to go to the original for further enlightenment. In the Campbell and Stanley notation:

> "An X will represent the exposure of a group to an experimental variable or event, the effects of which are to be measured; O will refer to some process of observation or measurement; the X's and O's in a given row are applied to the same specific persons. The left-to-right dimension represents the temporal order, and X's and O's vertical to one another are simultaneous." (p. 6)

Example

As an example of this notation system, consider the case where our objective is the delivery of an innovative health service. We have confidence that our health service is beneficial to the recipients, but since many of our colleagues and friends are skeptical, we decided to document the effects of our endeavors by measuring the initial level of disability, providing the health service, and once again measuring the level of disability. In our Campbell and Stanley notation, this strategy of service delivery and evaluation thereof might be depicted as follows:

$$O_1 \; X \; O_2$$

The first O_1 indicates the intitial measurement of disability, the X represents our intervention (delivery of innovative health care), and O_2 represents the posttreatment assessment of disability level.

Supose that at O_1 we measure the extent of disability exhibited by 100 clients, and find that the average level of disability is 70. At the second measurement occasion (O_2), after delivery of our innovative health program (the X, or **independent variable**), we discover that the average level of disability for the 80 clients remaining for the duration of our innovative program is 40. Eureka! A decrease in average disability from 70 to 40 (on a scale with 0 to 100 range) verifies our belief in the efficacy of our innovative health program. Or so we think initially.

But then one of our staff suggests that we should test the statistical significance of the outcome. Therefore, we arrange for the venerable Dr. Cracker from State Teacher's University to advise us on statistical manipulations of our little numbers. Dr. Cracker arrives, examines our little numbers, reduces us to trembling blobs of protoplasm with the comtemptuous remark that we were fools to contemplate the study without first consulting **him**, and then dons the robe of savior by assuring us that he can save the day by application of a correlated samples t-test. Dr. Cracker returns to State Teacher's University and forthwith there arrives by mail a computer printout which tells us, among other things, that the variances were homogeneous, there were 79 degrees of freedom, the obtained t-value was 4.63, and that a pre-post difference as large as we obtained (i.e., 70 versus 40) could be expected by chance less than once in a thousand occasions.

Elated, we rush to press with our findings, endeavoring to share with the literate world vindication of our belief in the efficacy of our innovative treatment program. We draft a manuscript, fire it off to the journal editor special delivery, and confidently await notification of acceptance of our article for publication.

When the manuscript arrives at the office of the journal editor, however, events do not transpire exactly as we had expected. The editor is not familiar with the statistical test that we report, so he asks a methodologically talented friend of his, Don Design, to review and comment upon the technical adequacy of our study. The editor, upon receipt of Don Design's comments, forwarded them to us. Our original manuscript had been 11 pages, it is now 33 pages. Don Design wrote twice as much as we. How could we have erred so bountifully? The sections below convey some of the ways in which Don Design edified us.

LESSONS OF DON DESIGN

Don Design commended our efforts of attempting to develop, implement, and evaluate an innovative health services program, and he was impressed that we took care to conduct a statistical test, but he cautioned us that statistical inference is not tantamount to causal inference. Since our primary objective was clarification of the efficacy of our innovative health program, our proper concern was with the veracity of the causal statement: Our innovative health program improves health. Thus, although we were correct in applying a statistical test, and even though we applied the correct test under the circumstances (thanks to the sage advice of Dr. Cracker), we failed to consider that a good experimental design is the necessary foundation upon which statistical edifices may be erected. Without an adequate underlying experimental design, even the most sophisticated tests of statistical significance are meaningless.

Purposes of Experimental Design

The overarching purpose of any experimental design (and there are scores of experimental designs) is to assist in establishing causal relationships. Specifically, an experimental design can serve in two important ways:

1) As a superordinate conceptual template to guide:
   - A. the assignment of experimental units to treatment conditions (e.g., which people receive what type of health service)
   - B. the manipulation of independent variables (The independent variable (IV) is controlled by the investigator. The impact of the IV is observed upon the dependent variable (DV), in our example, some index of health.)
   - C. the collection of data (measurements and/or observations).
2) As a means to the elimination of plausible alternative explanations.

The quality of an experimental design may be judged on the basis of how much experimental control it affords the investigator. Experimental control in this context means the extent to which plausible rival hypotheses (or extraneous variables) can be controlled or ruled out. Experimental designs are commonly evaluated upon the bases of how adequately they protect against threats to internal validity and threats to external validity.

Internal and external validity. An investigation possesses internal validity to the extent that we can be assured that an observed effect was really caused by the experimental manipulation (the X, or in our case the innovative health service) and not by some extraneous variable. External validity concerns the extent to which we may by confident that the same results would be obtained with a different population (population external validity) and/or in a different setting (ecological external validity). Generally, external validity is concerned with the generalizability of a particular obtained result.

Don Design found few faults with our investigation in the area of external validity, but he criticized us extensively on the basis of internal validity (actually, the lack of internal validity). Don Design enumerated several threats to internal validity against which our design provided little or no protection. That is, several plausible alternative explanations of our finding could not be ruled out, and hence we could not be certain of what caused the obtained results.

With unmitigated temerity, Don Design informed us that we had not even used an experimental design! Rather, our investigative paradigm might be characterized more accurately as a "pre-experimental" design, or even as a "pseudo-experimental" design.

Our critic asked: If you were interested in assessing the efficacy of your innovative health program, against what standard should it be compared? Our program is innovative, we thought (and so did our critic), so it should be evaluated relative to traditionally available services. Then why not employ an experimental design that would both permit a comparison with traditionally avaiable services and provide protection from threats to internal validity?

"Pre-experimental" and "True" Experimental Designs

According to Don Design, we had employed a pre-experimental design which he called the One-Group Pretest-Posttest Design. How much better, lamented our critic, if we had utilized a Pretest-Posttest Control Group Design, a true experimental design which protects against all major threats to internal validity under most circumstances. Consider the tow designs:

Our Design: One-Group Pretest—Posttest	**Recommended Design: Pretest—Posttest Control Group Design**
O_1 X O_2	R O_1 X O_2
	R O_3 Z O_4

In the recommended design, we would have assigned randomly half our participants to the group receiving the innovative treatment (X), and the other half to the (control or comparison) group receiving the traditional health services (Z). Random assignment in this design (signified by the R's) is critically important. Random assignment (meaning that each participant has an equal probability of being assigned to innovative or traditional services conditions) assures that, over the long-run, pretreatment equality of the groups on all relevant variables will be attained. If we had utilized this recommended design, instead of the pre-experimental design, we would have assigned each client randomly to one of the two groups, collected our pretreatment assessments of extent of disability, exposed one group to the innovative treatment and the other to the traditional treatment, and then taken our posttreatment assessments.

Supose that we had used the recommended design and observed that the two groups (each composed of 50 clients) exhibited comparable pretreatment average disabilities of 70, but that after treatment the 40 clients in the innovative group exhibited a mean disability of 40, while the mean for the 40 clients that received traditional services was 70.

What can we now conclude? That is, is there a causal relationship between treatment and disabiltiy? It appears that exposure to our innovative treatment produces markedly improved health in the sense of reduced disability, but that exposure to traditional services produces no change in health, vis a vis disability. Once again we consult with Dr. Cracker. This time he advises us to use either an independent samples t-test, or a single classification analysis of covariance. We choose the latter, since it sounds more recondite, and are informed, among other things, that an adjusted mean difference as large as the one obtained could be expected to occur by chance less than one time in a thousand occasions.

But isn't that what Dr. Cracker concluded about our initial results, using the inadequate pre-experimental design? Yes, but remember two points discussed earlier: 1) Statistical inference is not tantamount to causal inference, and 2) Some designs provide greater protection from threats to internal validity. Let us consider then, the extent to which these tow hypothetical sets of results (i.e., from the One-Group Pretest-Posttest Design on the one hand, and the Pretest-Posttest Control Group Design on the other) are protected from plausible alternative explanations by their respective designs.

Protection from Threats to Interal Validity

For purposes of exposition, our two designs will be referred to as Design Number 1 and Design Number 2 (as indicated below), as we consider several threats to internal validity.

Design Number 1	**Design Number 2**
O_1 X O_2	R O_1 X O_2
	R O_3 Z O_4

History. History refers to any event occurring between O_1 and O_2, external to the participants, which also could have produced the observed change (i.e., from 70 to 40). For example, consider that the pretreatment measurement occurred on March 1 and the posttest on May 31. Is it possible that the changes observed could be attributed not to the treatment received (X), but rather to the arrival of the spring season and the anticipated cheer and relaxation of approaching summer? Notice that in Design Number 1 this plausible alternative explanation cannot be ruled out. This is not the case in Design Number 2, however. If the approach of spring produced the reduction in mean disability from 70 to 40 for the innovative treatment group, then there should have been a similar reduction for the clients **randomly assigned** to the traditional services group. Thus, Design Number 2 emerges as decidedly superior to its counterpart, Design Number 1, because the former affords us protection from history as a threat to internal validity.

Maturation. Maturation refers to processes within the organism, occurring as a function of the passage of time, which might account for obtained results. Maturation is used in a rather broad sense to designate all those biological or psychological processes which vary systematically with the passage of time. In our example, a maturational alternative explanation for obtained results might be that of spontaneous remission. In using Design Number 1 a critic might ask us: How can you be sure that it wasn't merely the passage of time that produced the mean change from 70 to 40? Design Number 2, however, protects against this alternative explanation, since spontaneous remission should have occured equally for both groups.

Testing. The effect of testing (pretreatment measurement) may itself be confounded with the effect of the treatment program in Design Number 1. That is, the obtained change in disability may be caused by the pretest, and not by the treatment. For example, administration of the pretreatment tests and questionnaires may sensitize the clients to their condition and serve as a stimulus to change. Once again, however, Design Number 2 is not susceptible to this criticism, since both groups are pretested and therefore should exhibit equal reactions if the pretest is a plausible alternative explanation for obtained results. In cases where the pretest is clearly reactive, it can be omitted entirely and a different experimental design can be implemented. Alternatively, the investigator might strive to employ nonobtrusive measurements (Webb, Campbell, Schwartz, and Sechrest, 1966).

Instrumentation. This term refers to a change in the measuring instrument between pretest and postest which might account for any observed difference. Supose that pretreatment assessments had been made by Technician A, but that posttrestment assessments of disability were made by Technician B. If this had been the case, then an O_1-O_2 difference in Design Number 1 might be attributable to a change in instrumentation, but Design Number 2 protects aginst this threat, since the impact of the "instrument decay" variable should be manifest equally in both groups. Perhaps the best arrangement would entail using the same Technican for both pre- and posttreatment assessments, and have the Technician work blindly, that is, unaware of which treatment a particular client received. This procedure would safeguard against subtle experimenter bias effects, where the experimenter's biases are subconsciously reflected in the data due to selective perception. For example, if the Technician is convinced that the innovative treatment was superior, this bias might enter, albeit subtly and unconsiciously, into interpretation of client responses.

Statistical regression, or regression toward the mean, is a ubiquitous phenomenon where groups are selected on the basis of their extremity and there is imperfect test-retest reliability. An "elementary and old-fashioned exposition" of statistical regression is presented in Campbell and Stanley (1966; pp. 10-12), and more advanced treatments are provided in Campbell and Erlebacher (1970a; 1970b). For our purposes, it is most important to recognize that whenever a group is selected on the basis of extreme scores (in our example, a high degree of disability) and then is remeasured, the group will regress toward the overall population mean on the second testing occasion, if the test-retest correlation is less than unity. The amount of statistical regression exhibited is a function of the test-retest correlation of the instrument in question.

Consider the hypothetical case where an academic achievement test was administered to all 10th graders in school system B. The top 50 scorers (gifted youth) and the bottom 50 scorers (scofflaws) were identified and enrolled in a 6 month course designed to increase motivation to succeed academically. It was hypothesized that this treatment would produce great scientists and playwrights among the gifted youth, and transform the scofflaws into mediocre achievers. Actually, there were absolutely no effects produced by the motivation course. After the course, the achievement test was administered again. The 6 month test-retest correlation of the achievement measure was .50, and the results are depicted below.

(figure 1)

Notice that both of the extreme groups regressed half the distance to the overall mean although there was **no** true treatment effect (the test-retest correlation was 0.50). In our hypothetical Design Number 1 investigation, where we studies a single extreme group, we could expect some degree of improvement from O_1 to O_2 on the basis of statistical regression alone (even with absolutely no effect attributable to the treatment) if the test-retest reiability of the measuring instrument was less then unity. Hence, in Design Number 1 a plausible alternative explanation is that the observed reduction in degree of handicap was due to statistical regression. With Design Number 2, on the other hand, statistical regression is not a plausible alternative explanation. If the innovative treatment group was affected by statistical regression, then the traditional services group should have been affected similarly, since they were comparably extreme on pretest scores. (Do you begin to see the value of randomized assignment?)

Mortality. This term refers to loss of participants over the duration of an investigation, and is often as much a problem for Design Number 1 as for Design Number 2, where it sometimes may be the case that experimental and comparison groups reflect differential attrition (not the case in our hypothetical study, since 80% of the participants in both groups were posttested).

Summary. Clearly, Design Number 2 (a true experimental design) provided greater protection from threats to internal validity than did Design Number 1 (a pre-experimental design). Since the most imporatant objective that an experimental design can accomplish is to control for plausible alternative explanations that might compete with the hypothesized cause-effect relationships, we are persuaded by Don Design that we profitably could have utilized a true experimental design.

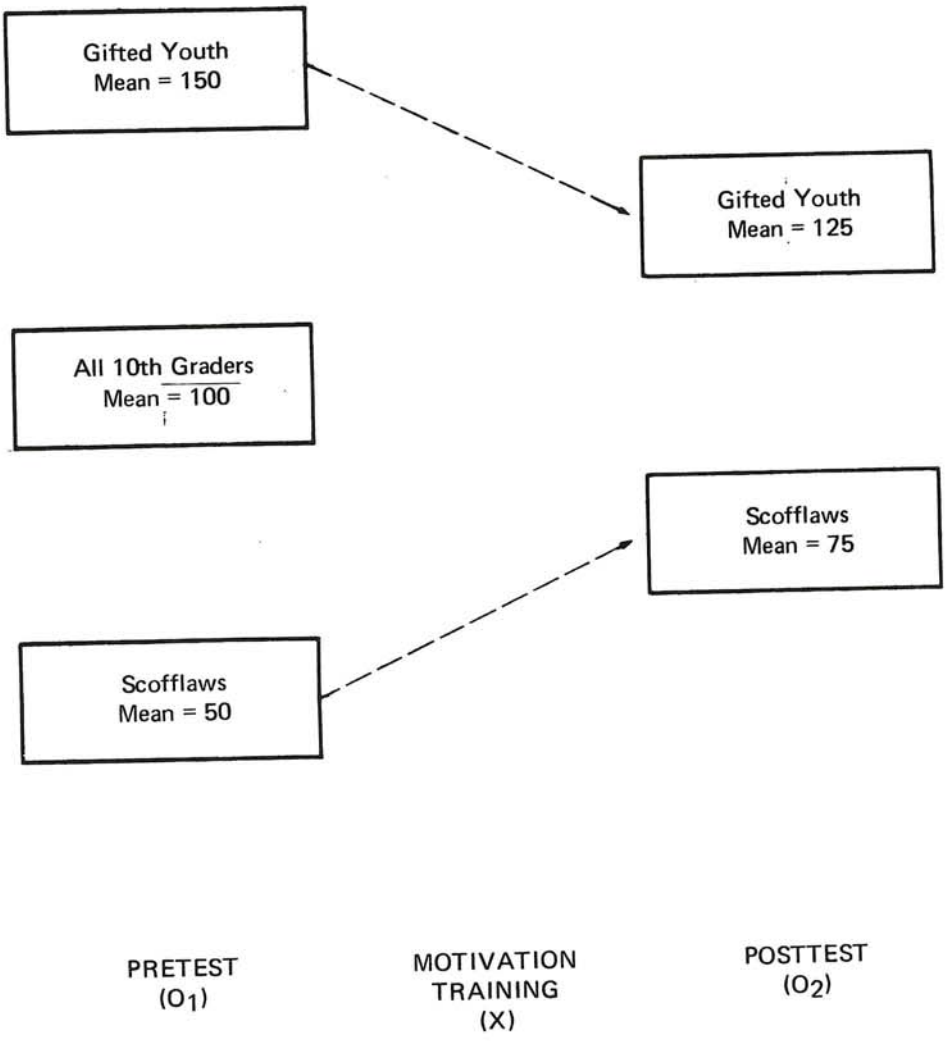

Figure 1. Illustration of how statistical regression (toward the overall mean) would affect posttest scores of two extreme groups with test-retest reliability of 0.50 and absolutely no true treatment (motivation training) effect.

OTHER CONCERNS

Thus ends the saga of our hypothetical health program, Don Design, and Dr. Cracker from States Teacher's University. For the reader who takes an interest in little numbers, Research Avenue, Statistics Boulevard, evaluation research, and experimental design, the remainder of this paper is devoted to a short annotated bibliography of recommended reading.

Experimental and Quaxi-Experimental Designs for Research: Don Campbell and Julian Stanley (1966).

This little treatise (only 84 pages) is a masterpiece. Originally published in 1963, it was characterized recently in this manner: "If there is a Bible for evaluation, the Scriptures have been written by Campbell and Stanley" (Rossi & Wright, 1977, p. 13).

Campbell and Stanley is not easy reading, but if you persevere you will be rewarded well. You will learn that the Pretest-Posttest Control Group Design (our Design Number 2 above) is just one of several true experimental designs, all of which must be evaluated (in terms of how well they protect against validity threats) in the particular context in which they are used. (In other words, even a true experimental design can be botched.)

This treatise also discusses several representatives of a family of investigative paradigms called quasi-experimental designs, which can be used in many circumstances where it is impractical to utilize a true experimental design. For example, in situations where randomized assignment to experimental and treatment conditions are not possible, it may be possible to use some variant of the Multiple Time-Series Design.

Design and Analysis of Time-Series Experiments: Glass, Wilson, and Gottman (1975).

Campbell and Stanley excited everybody with the potential of time-series designs. Here is an entire book on the topic. Read and enjoy. The first four chapters assume no mathematical sophistication.

Reading Statistics and Research: Huck, Cormier, & Bounds (1974).

Chapters 11 to 14 provide a splendid introduction to experimental design. Heavily influenced by Campbell and Stanley, yet much easier to read. The authors do a fine job of illustrating how simple true experimental designs serve as basic building blocks for more complex factorial designs. Well-worth the moderate purchase price.

Handbook of Evaluation Research: Struening and Guttentag, Eds. (1975).

A massive two-volume work (over 1,400 pages), the handbook ranges across most areas of evaluation: Policy and Strategy, Experimental Design, Development and Evaluation of Measures, Interview Methods, Analytic Methods, Politics and Values, and more. It also features chapters by noted authorities summarizing the state of the art of evaluation research within their respective substantive specialities. Some of the areas reviewed: Public Health Program, Mental Health Services, Early Intervention Projects, and Residential Treatment Programs for Disturbed Children.

Evaluation Studies Annual Review: Glass, Ed. (1976).

This is Volume 1 of an anticipated annual collection of the best publications in the general area of evaluation research. Contains an informative and penetrating review of the *Handbook,* discussed above.

Evaluation Quarterly

A brand new journal devoted to evaluation. No. 1, Vol. 1, was issued February, 1977. The journal, purportedly, will concern itself with: 1) "articles either that make significant empirical contributions or that develop new research techniques in evaluation research", 2) "papers that integrate findings and perspectives", 3) "brief reports of research efforts and investigations in progress", and 4) "brief 'Craft Reports' with a 'how-to-do-it' flavor".

Social Experimentation: A Method for Planning and Evaluating Social Intervention: Riecken and Boruch, Eds. (1974).

Long-awaited and somewhat disappointing (overly-edited) this major work by a committee of the Social Science Council reflects the contribution of Donald Campbell. As the title accurately implies, the emphasis throughout is on **experimental** program evaluation. Addresses many issues: design and analysis, measurement, execution and management, political considerations, human values. Concludes with a useful chapter outlining illustrative controlled experiments for planning and evaluating social programs.

Evaluative Research: Suchman (1967).

The first comprehensive statement on evaluation research. Chiefly of interest historically. Reflects methodological naviete.

Methods for Experimental Social Innovation: Fairweather (1967).

A comprehensive work on the feasibility of experimentally assessing the efficacy of **organized** social innovations. Considers social innovations as social sub-systems, or alternatives to pre-existing social organizations. Verbose, but appealing.

Reading in Evaluation Research: Caro, Ed. (1971).

Probably still the best book of readings in the area. Contains Campbell's (1969) *American Psychologist* "Reforms As Experiments" article, plus an eye-opening article published in *Social Forces* in 1935. Also contains useful statements by Scriven, Brooks, Rossi, Weiss, Greenberg, Weiss and Rein, Evans, and others. A total of 31 articles.

Evaluating Social Programs: Rossi & Williams, Eds. (1972).

Another good book of reading. Tripartite organization: Theory, practice, and politics. Looks at difference kinds of social programs and their evaluation: Compensatory education, federal manpower programs, income maintenance experiments; others.

Evaluating Action Programs: Weiss, C. H., Ed. (1972).

Another book of readings. Available in inexpensive paperback. Good statements by Weiss, Alkin, Rossi, Anderson and Sherwood, others. Some overlap with Caro (1971) articles. Twenty-one articles.

Federal Evaluation Policy: Wholey, et. al. (1970).

A view from the perspective of the "consumers" of evaluation endeavors. One good chapter on methodology.

Educational Evaluation: Stufflebeam, et al (1971).

Reads as though written by seven authors, which it was. Unpersuasive and simplistic in places, it is nevertheless worthwhile because of the Stufflebeam influence evident throughout.

Educational Evaluation: Theory and Practice: Worthen and Sanders, Eds. (1973).

Approaches **educational** evaluation as disciplined inquiry, distinct from research. Each major section is introduced by editorial comments, which are quite penetrating. Possibly the best work on educational evaluation.

Evaluation of Behavioral Programs in Community, Residential, and School Settings: Davidson, et. al, Eds. (1974).

The central topic is program evaluation in social and health programs. Chapter content is variable, ranging from a discussion of the complementary use of single-subject designs and traditional comparison group experimental designs to evaluations of juvenile correction programs, mental health programs for the aged, community-based psychiattic services, etc. Concludes with a chapter on evaluation of program evaluations.

REFERENCES

Campbell, D. T. Reforms as experiments. In E. L. Struening and M. Gutentag (Eds.), *Handbook of evaluation research* (Vol. 1), Beverly Hills: Sage, 1975, 71-100.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Eds.), *Disadvantaged Child* (Vol. 3). New York: Brunner/Mazel, 1970, 185-210. (a)

Campbell, D. T., & Erlebacher, A. Reply to the replies. In J. Hellmuth (Ed.), *Disadvantaged Child* (Vol. 3). New York: Brunner/Mazel, 1970, 221-225. (b)

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.

Caro, F. G. *Readings in Evaluation Research.* New York: Russell Sage, 1971.

Davidson, P. O., Clark, F. W., & Hamerlynck, L. A. (Eds.) *Evaluation of Behavioral Programs in Community, Residential and School Settings.* Champaign, Ill.: Research Press, 1974.

Fairweather, G. W. *Methods for Experimental Social Innovation.* New York: John Wiley and Sons, 1967.

Ferman, L. A. Some perspectives on evaluating social welfare programs. *Annals of the American Academy of Political and Social Science,* 1969, **385**, 143-156.

Glass, G. V. (Ed.) *Evaluation studies review annual* (Vol. 1). Beverly Hills; Sage, 1976.

Glass, G. V., Wilson, U. L., & Gottman, J. M. *Design and analysis of time-series experiments.* Boulder: Colorado Associated University Press, 1975.

Huck, S. W., Cormier, W. H., & Bounds, W. G., Jr. *Reading statistics and research.* New York: Harper & Row, 1974.

Riecken, H. W., & Boruch, R. F. (Eds.) *Social experimentation: A method for planning and evaluating social intervention.* New York: Academic, 1974.

Rossi, P. H., and Wiliams, W. (Eds.) *Evaluating Social Programs.* New York: Seminar, 1972.

Rossi, P. H., & Wright, S. R. Evaluation research: An assessment of theory, practice, and politics. *Evaluation Quarterly,* 1977, **1**, 5-52.

Struening, E. L., & Guttentag, M. *Handbook of evaluation research.* Beverly Hills: Sage, 1975.

Stufflebeam, D. I., et al. *Eduational Evaluation.* Itasca, Illinois: F. E. Peacock, 1971.

Suchman, E. A. *Evaluative Research.* New York: Russell Sage, 1967.

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures: Nonreative research in the social sciences.* Chicago: Rand McNally, 1966.

Weiss, C. H. (Ed.) *Evaluating Action Programs.* Boston: Allyn and Bacon, 1972.

Wholey, J. S., Scanlon, J. W., Duffy, H. G., Fukumoto, J. S., and Vogt, L. M. *Federal Evaluation Policy.* Washington, D. C.: The Urban Institute, 1970.

Worthen, B. R., and Sanders, J. R. (Eds.) *Educational Evaluation: Theory and Practice.* Worthington, Ohio: Charles A. Jones, 1973.