THE PITFALLS OF MANPOWER EXPERIMENTATION

Gus W. Haggstrom

RAND Corporation

Prepared for:

Defense Advanced Research Projects Agency

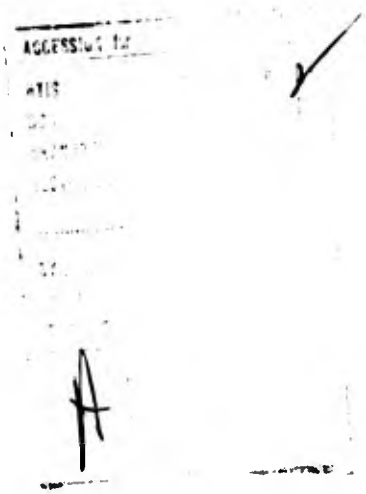
April 1975

THE PITFALLS OF MANPOWER EXPERIMENTATION

Gus W. Haggstrom

April 1975

P-5449



The Rand Paper Series

Papers are issued by The Rand Corporation as a service to its professional staff. Their purpose is to facilitate the exchange of ideas among those who share the author's research interests; Papers are not reports prepared in fulfillment of Rand's contracts or grants. Views expressed in a Paper are the author's own, and are not necessarily shared by Rand or its research sponsors.

The Rand Corporation
Santa Monica, California 90406

THE PITFALLS OF MANPOWER EXPERIMENTATION

by Gus W. Haggstrom[1]

The Rand Corporation, Santa Monica, Calif.

April 1975

SUMMARY

Controlled experiments to test new military personnel policies on a small scale before they are implemented have many advantages over other techniques for evaluating the potential effects of these policies. However, case studies of recent experiments in the Army and Air Reserve Forces demonstrate the hazards of conducting such studies unless precautions are taken to assure the validity of the experimental results. The Army's experiment to test the feasibility of shortening the term of enlistment in the reserves provides a particularly good example of how not to conduct a field study of this type. The lessons learned from this and other experiments underscore the necessity of following certain guidelines in planning and conducting such tests in the future.

---

[1]The research described in this paper was primarily supported by the Defense Advanced Research Projects Agency under contract No. DAHC15-73-C-0181.

This paper was prepared for presentation at the Conference on Survey Alternatives, Santa Fe, New Mexico, April 22-24, 1975. The views expressed in this paper are the author's own and are not necessarily shared by The Rand Corporation or its research sponsors.

INTRODUCTION

On June 1, 1973, the Air Reserve Forces began an experiment to test whether reducing the term of enlistment for nonprior servicemen would have a substantial effect upon recruiting. Prior to the experiment, all male nonprior service enlistees had to enlist for a period of six years. During the experiment, a few carefully selected units were permitted to offer potential recruits a "3x3" enlistment option--three years of regular reserve duty followed by three years in the Individual Ready Reserve (IRR). Since participation in the IRR does not entail drill attendance, this option effectively reduced the enlistment period to three years. Certain other units were permitted to offer a "4x2" option--four years of regular reserve duty followed by two years in the IRR.

A month after the Air Force experiment began, the Army Reserve Components undertook a similar experiment except that, instead of offering the options in just a few reserve units, the Army offered the 3x3 option in all reserve units in 16 states and the 4x2 option in 12 other states and the District of Columbia.

The results of these experiments will be summarized later in the paper. In brief, the shortened enlistment options proved to be far less effective in attracting new recruits than the military had expected. In particular, the experimental evidence indicates that adopting the 3x3 option would not attract a sufficient number of additional recruits to offset the man-year losses that would result later on.

The primary purpose of this study is not to present the experimental findings on the attractiveness of shorter enlistment tours but to report some of the other lessons learned from these experiments. Whereas the Air Force test satisfied most of the criteria usually prescribed for carrying out experiments of this type, the Army test deserves to be cited as an example of how not to conduct an experiment. The two tests taken together provide excellent case studies to illustrate the benefits and hazards of manpower experimentation and demonstrate the need for establishing guidelines for future experimentation.

In theory, the notion of conducting controlled field studies to test new personnel policies on a small scale before they are implemented has considerable merit. The advantages of such tests relative to other information-gathering techniques are reviewed in the next section. Then some of the shortcomings of manpower experiments will be considered, and guidelines for conducting such experiments will be given. A technical appendix outlines the statistical theory for designing experiments and treats some of the theoretical aspects involved in using linear models to analyze these experiments. Finally, the experiments mentioned above will be used as case studies to illustrate the pitfalls of manpower experimentation and how to avoid them.

WHY EXPERIMENT?

In deciding whether to implement a new personnel policy, military policymakers are usually content to rely upon expert opinion, anecdotal evidence, sample surveys, or analyses based on nonexperimental data for assessments of the potential effects. The primary reason for undertaking controlled experiments to test new policies is to gather more reliable information about the effectiveness of those policies than can be obtained in other ways. Whereas the other methods rely on indirect evidence, controlled experiments attempt to evaluate the policies by putting them into practice in such a way as to permit valid estimates and more precise comparisons of their effects.

It should be clear from the foregoing that the type of experiment under consideration here usually consists of administering one or more treatments (or policies or programs or options) to subgroups of people or other units of observation. The purpose is to compare the responses of the experimental units receiving the different treatments in order to estimate the effects of the treatments. By a "controlled" experiment is meant one in which steps are taken to hold constant other factors that might affect the responses of the experiments units. Usually such experiments are characterized by the use of randomly chosen "treatment groups" and a "control group" of persons (or military units or training classes or locations) that do not receive the treatments under study but are observed in the same manner as the treated individuals during the course of the experiment.

As a simple example, consider a hypothetical experiment to evaluate a new training program for trainees in a certain military occupational specialty. To see how well this program works relative to the usual training program, one could choose a class of students at random from the next cohort of entering trainees and a corresponding control group who would receive identical treatment in all respects except for the method of training. The responses of interest might be the trainees' performances on some test to measure their proficiency in carrying out their duties.

A similar experiment could be devised to test the effectiveness of a particular type of advertising campaign upon recruiting. Suppose there are a number of relatively homogeneous recruiting stations (or locations) that could be used in the study. Then the experiment could be carried out by choosing a certain subset of the recruiting stations at random and subjecting them to the advertising campaign, using the other stations as a control group. A comparison of the enlistment rates for the two groups would yield an estimate of the effect of the advertising campaign.

For reasons to be explained later, simple experiments of this type are rarely carried out in this manner. Ordinarily, the experimental situation is exploited to derive other information of relevance about the process under study. For example, in addition to seeing how well students trained under a new program perform relative to the control group, one would probably want to determine if the new program works better for training students who have low Armed Forces Qualification Test (AFQT) scores. These considerations and others would probably lead to assigning the trainees to the treatment and control groups in other than the completely random manner indicated above.

As a more realistic example of a complex manpower experiment, consider setting up a test of various types of recruiting effort and incentives by experimenting with (a) different types of advertising, (b) different levels of recruiter activity, and (c) various recruiting incentives. Here, the observational units for the experiment might consist of geographical subdivisions or military units, and the responses of interest might be the number of enlistments at the various locations or some enlistment rate that allows for differences in population sizes. As this example illustrates, controlled experiments can be used to test the relative effectiveness of several alternative proposals simultaneously.

Some alternatives to controlled experiments for gathering information about new personnel programs are: (1) expert opinion, (2) anecdotal evidence, (3) sample surveys, and (4) analyses based on managerial records and other nonexperimental data. Each of these methods will be discussed briefly below.

Expert opinion. Military policymakers have often relied heavily on expert opinion in devising and implementing personnel programs. The studies prepared for the Gates Commission provided expert guidance, some of which was based upon analyses of nonexperimental data, for programs that might be undertaken during the transition to an all-volunteer force.[1] More commonly, the solicitation of expert opinion means asking those that are most closely associated with the process under study how well they think the policy will work. Thus, instructors, training school commandants, and educational experts may be asked to evaluate a proposed change in training techniques. Recruiters and their supervisors may be questioned about the possible effects of shortening the enlistment tour or offering an enlistment incentive of a certain type.

To illustrate how misleading expert opinion can be, a survey of Air National Guard recruiters indicated that shortening the tour of enlistment in the ANG would result in a three-fold increase in enlistments. Yet, when the variable tour experiment was conducted to test this contention, there were virtually no differences in the overall enlistment rates of the ANG units that offered shorter enlistment tours and those that did not.

Expert opinion can be wrong, misleading, and biased by selfish motives or group pressure. One would hope that this is the exception, not the rule, but the experts may lack the high-quality information necessary for making clear-cut evaluations of new policies. In the absence of controlled experiments in certain areas, their expertise

---

[1] Studies Prepared for the President's Commission on an All-Volunteer Armed Force, Vols. I & II, Washington, D.C., November 1970.

may be limited to educated guesses. This is not to say that controlled experiments can supplant expert opinion. Indeed, the guidance of those who are most knowledgeable about the problem under investigation is essential to all stages of controlled experimentation. Experts are needed to help formulate the policies, to specify the criteria by which the program's performance should be gauged, to help design the study, to pinpoint the variables that may have a significant bearing on the process under study, and to bring relevant textbook theory and past research to bear on the conduct and analysis of the experiment.

Anecdotal Evidence. This type of evidence consists of informal comments by people who report their observations and criticisms of the system under study. Such evidence might be gathered through a "Suggestion Box," by soliciting comments at the end of an opinion survey, or through casual conversations. In medical research the carefully documented case studies reported in medical journals are an important type of anecdotal evidence leading to the development of lists of symptoms associated with certain diseases and reports of successful treatments.

Anecdotal evidence, coming from a variety of sources of unknown reliability and with unknown biases, defies statistical analysis. Students who drop by after class to provide feedback to their instructors may present a far rosier picture of the course than those who don't. Letters to congressmen about the treatment of recruits in basic training may overstate the need for change. In those cases where anecdotal data are systematically collected from all participants in the program or from a randomly selected sample, the informal opinions may still be misleading. The people who work within a particular system day after day become stereotyped in their thinking about their activities and fail to consider alternatives that would occur to outsiders. They may dismiss the possibility of improving their system, or they may react defensively to observers, especially when their daily routines are threatened. On the other hand, the discovery of important innovations often stems from anecdotal evidence, and solicitation of this type of information from participants in a controlled experiment may enhance the discovery of new

knowledge and lead to significant improvements in the program.

Sample Surveys. Sample surveys are more systematic than anecdotal evidence for gathering opinions from members of the population that might be affected by the proposed program. As in the Gilbert Youth Surveys,[1] they are commonly conducted by having respondents fill out questionnaires. The respondents are usually selected using well-established sampling techniques to assure that they are representative of the population under study. The statistical theory behind sample surveys is well-developed,[2] and practice has become more and more consistent with theory. Sample surveys are versatile tools for policy analysis in that many items can be included on the questionnaire to elicit the respondents' opinions and possible reactions to various program options that may be under consideration. By gathering information on the personal characteristics of the respondents, sample surveys can indicate how various subpopulations might respond to different options, thus permitting program managers to tailor their programs to have greater appeal to individuals having certain characteristics. Still another appealing attribute of sample surveys is that they are relatively cheap information-gathering devices, particularly if one ignores the time spent by the respondents in filling out the questionnnaires.

Insofar as evaluating proposed personnel policies is concerned, the major shortcoming of the sample survey is not due to the statistical unreliability of the sampling process, but to the factual unreliability of the respondents and to uncertainties associated with predicting people's future reactions to policy changes based upon their quick responses to hypothetical questions. No matter how large the sample size, if some of the respondents deliberately or unconsciously distort their answers in

---

[1] John Goral and Andrea Lipowitz, "Attitudes of Youth toward Military Service in the All-Volunteer Force: Results of Gilbert Youth Surveys, May 1971 to November 1973," Manpower Research and Data Analysis Center, MARDAC MR 75-1, September 1974.

[2] See W. Edward Deming and Alan Stuart, "Sample Surveys," International Encyclopedia of the Social Sciences, Vol. 13, The Macmillan Company & The Free Press, Riverside, New Jersey, 1968, pp. 595-616, and the references cited there.

the same way, the results of the survey and deductions based on them will be subject to error. And even if the respondents are perfectly truthful, the artificiality of the survey technique is such that responses to hypothetical policy-related questions may not be indicative of the actual changes that would result if the policies were put into effect.

The type of survey that is especially prone to this kind of distortion is one in which the respondent is asked what his present plans are and then is asked how his plans would change if a certain policy were enacted. The ostensible purpose of such a survey is to pool the individual responses to predict what would happen if the policy were to be implemented. However, the lack of realism associated with responding to questions of this type leads to a certain amount of gamesmanship on the part of the respondents. No matter what degree of anonymity is assured the respondents, some will distort their answers to conform to what they feel are "normal" or desirable responses. Others will try to answer in such a way as to bring about changes that they feel would benefit themselves or their friends. Still others may just try to confound the statisticians. Even if the respondent tries to answer the questions conscientiously, his quick answers to a large number of questions may not correspond to his responses after giving the questions more consideration. Moreover, the questions themselves may not be clear to the respondents, and a trivial rewording of a question or a reordering of a sequence of questions may elicit substantially different results.

As an illustration of the hazards of such methods, a survey of U.S. Army Reservists in late 1971 led to the inference that, by merely offering the reservists the option of carrying $15,000 of Servicemen's Group Life Insurance at $3 or $4 per month, reenlistment rates in the USAR would rise from 15 percent before the policy was enacted to 23 percent afterwards.[1] This incentive was later enacted by Congress. But since

---

[1]W. L. Clement et al., Maintenance of Reserve Components in a Volunteer Environment, Research Analysis Corporation, RAC R-148, McLean, Virginia, November 1972.

$15,000 of SGLI at those prices hardly qualifies as a bargain, it seems safe to say that this option had at most a negligible effect upon reenlistment rates in the reserves.

Analysis of Nonexperimental Data. Analyses of managerial records, time series data, personnel files, and other nonexperimental data are also used to estimate the potential effects of policy changes. Here the idea is to exploit the natural variation in policies that various subgroups of people have experienced in the past to surmise how outcomes of interest (reenlistment rates, measures of productivity, job satisfaction indices, etc.) are related to various inputs (modes of compensation, manning levels, use of job performance aids, training techniques, etc.). The methodology usually involves fitting equations to historical data using the output measures as dependent variables and input measures as independent variables. The rationale is based on the shaky premise that these equations can then be used to estimate how manipulating the values of the independent variables through policy changes will affect the output measures.

There are many hazards associated with this approach. The formulas derived from historical data may be distorted because of incorrect specification of the functional form of the fitted equations, the exclusion of important missing variables, or biases in the data resulting from inaccurate, incomplete, or contrived historical records. Moreover, the method depends upon the existence of a clear-cut cause-and-effect relationship between the independent variables and the measures of output. However, it is usually impossible to ascertain to what extent the observed relationship results from causal links among the variables and to what extent it merely reflects the way the variables are associated with each other because of their mutual dependence on latent variables, errors of measurement, and random fluctuations.[1] Unless the causal links

[1] For theoretical discussions of the hazards of inferring causality in regression situations, see George E. P. Box, "Use and Abuse of Regression," Technometrics, Vol. 8, No. 4, November 1966, pp. 625-629, and W. G. Cochran, "The Planning of Observational Studies of Human Populations," The Journal of the Royal Statistical Society, Series A, Vol. 128, 1965, pp. 234-265.

are clear and the equations are correctly specified, changing the levels of the independent variables to achieve the results promised by the fitted equations amounts to wishful thinking. If the estimates of the policy effects involve extrapolations beyond the range of the historical data, further hazards come into play. A formula that provides an excellent approximation of a relationship among variables over one region may fit very poorly over another region. Just as adding too much fertilizer can kill the plants, changing the input variables in a personnel system beyond their usual ranges may drastically alter the input-output relationship.

The Advantages of Controlled Experiments. Each of the alternatives above uses indirect evidence to evaluate the potential effects of personnel policies, whereas in controlled experiments the policies are put into practice on a trial basis to provide direct evidence of their worth. As the prominent statistician G.E.P. Box wrote, "To find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)."[1] Of course, there are certain types of human experiments that cannot be conducted, and others are not feasible for various reasons. But in those cases where controlled experiments can be conducted, they offer more credible information than the alternatives, primarily because the policies are tested in practice and not by opinion polls or other indirect methods.

A second reason for attaching more credibility to controlled experiments is that they can be carried out in such a way as to "control" the other factors that might otherwise tend to confound the cause-and-effect relationship between treatments and outputs. There are obvious reasons for attaching little credibility to an observational study of a new training program in which the experimental class consists of the first thirty trainees who volunteer for the special program. Even if the new program has no beneficial effect whatsoever, the volunteers may still outperform the others by a wide margin, perhaps because

---

[1] George E.P. Box, ibid., p. 629.

they are more highly motivated or have other special attributes. A controlled experiment with randomly chosen treatment groups and other safeguards would control this "selection bias," a worrisome feature of observational studies.

Another feature associated with testing new policies on a small scale is that, in the event that the policy proves ineffective, it can be modified or scrapped. Thus, controlled experiments can be used to minimize the negative effects of poorly conceived or non-productive policies.

Even if the program under test proves to be a roaring success, the experiment may turn up some "bugs" in the program that can be eliminated before the policies are implemented. The process of designing and conducting a controlled experiment can be a very enlightening experience for all concerned. The investigators and participants will learn a great deal about "what's going on," and this increased knowledge of the system will ordinarily lead to significant improvements in the programs under consideration, no matter how well-conceived the programs were initially. Moreover, the close monitoring of the experimental responses and the determination of performance criteria for the alternative options under test may lead to valuable quality control methods and measures of performance for the system once the new program is under way.

Controlled experiments, like clinical trials to test drugs, provide a fertile setting for making discoveries about the system under study that may lead to favorable returns far beyond the original goals of the experiment. For example, consider an experiment to evaluate the effectiveness of certain recruiting incentives in which the incentives are assigned to specially selected units and the other units are used as a control group. In the process of the carrying out the experiment, it will surely be discovered that certain units in the control group seem to be performing unusually well compared with other units having similar characteristics. Are these "outliers" attributable to outstanding recruiters, unusually productive recruiting techniques, peculiar

economic phenomena, or just unexplainable "randomness"? Questioning the recruiters and the commanders of the outstanding units may lead to a successful national recruiting campaign or an improved method of screening recruiters. A pattern of poor recruiting in college towns or high-income cities may indicate undesirable locations for new unit assignments. This example shows how data analysts can exploit the higher quality data available from experiments to isolate outlying individual performances and detect patterns among the variables that may have important policy implications.

The Shortcomings of Controlled Experiments. Although controlled experiments have certain clear advantages over other methods for gathering information on personnel systems, they also have certain drawbacks. Perhaps their main shortcoming is that they are more expensive and time-consuming than the other methods. It takes a lot of time to design an experiment properly, get everyone involved to agree on the particulars, conduct the experiment, and analyze the results. Hence, controlled experiments should not be undertaken unless the payoff in information justifies the time and effort required to carry them out.

For example, consider the survey result mentioned previously which led some to infer that offering SGLI at nominal rates to reservists would result in a substantial increase in reenlistment rates. Before Congress enacted this proposal in 1974, one might have been tempted to conduct an experiment to test the attractiveness of this incentive. However, it is doubtful that the experiment would be worth the trouble. Since the program would cost the military little, the risks associated with adopting the policy were minimal. Second, a test of the program would have to run for a long time to detect the small change in reenlistment rates that might result. The only apparent advantage to conducting a test of this type would be to assess the validity of conclusions drawn from sample surveys, which is an area that clearly needs more research.

Although controlled experiments can be expensive and time-consuming, many of the costs associated with experimentation are start-up costs that

the new programs would have to bear anyway. Efficiencies learned in putting the policy into practice on a trial basis may lead to reduced costs in implementing the program later, and the elimination of "bugs" may enhance the success of the new program.

Other shortcomings of controlled experiments for testing personnel policies derive from the fact that there are special problems associated with experiments in which the experimental subjects are people. First, there are the ethical considerations involved with "human engineering." If the treatments under test involve providing special privileges to selected groups of personnel, questions of inequity arise. Military leaders seem particularly sensitive to this aspect of experimentation. However, as in medical trials to test new drugs or different treatments, possible inequities associated with using different treatments for different groups of people must be weighed against the benefits to be achieved.[1] Given the magnitude of the costs associated with adopting ineffective personnel policies in the military and the foregone opportunities to invest these resources in other ways, the military has an obligation to evaluate new programs as carefully as possible before they are implemented. Thus, the decision not to experiment also involves ethical considerations.

A second difficulty associated with using human experiments to evaluate personnel policies is that people are notoriously poor experimental subjects, and human experiments are much more vulnerable to challenge than other experiments. Man is a complex, whimsical creature. His frailties, idiosyncrasies, and biases make it uniquely difficult to experiment on him, and having humans both as experimenters and as experimental subjects compounds the problems. Thus, many safeguards must be taken in conducting these experiments. And no matter how carefully they are conducted, experiments in the behavioral sciences lack the credibility of other scientific studies.

---

[1]For an interesting case study of the problems and ethical issues associated with clinical trials, see Paul Meier, "The Biggest Public Health Experiment Ever: The 1954 Field Trial of the Salk Poliomyelitis Vaccine," in Judith M. Tanur, ed., Statistics: A Guide to the Unknown, Holden-Day, Inc., San Francisco, 1972.

The reasons for widespread skepticism toward human experiments are well-founded. Many "controlled" experiments in the field of education that originally supported claims of remarkable advances in educational methodology were later refuted by more carefully conducted studies. Findings from medical experiments that presumably established the beneficial effects of certain medical treatments were later cast in doubt when placebos yielded the same effect.[1]

A common concern in human experiments is that the responses of the participants may be inflated by "Hawthorne effects," i.e., changes in the individuals' responses that are not attributable to the treatments under study but to other factors, usually psychological in nature, that may affect some people's behavior in an experimental setting.[2] If the subjects like the idea of participating in the experiment, or they sense that their behavior is being monitored closely, or they feel that the treatment will help them, they may respond unusually well even if the treatment under study has no beneficial effect whatsoever. If the experimental situation seems too unrealistic or the participants sense that experimental findings might benefit them or their friends, they may act abnormally to distort the findings and confound the analysts. Still another concern is that the experimenters themselves may take actions, perhaps inadvertently, that favor some experimental group or make a particular treatment look good.

Experimenters can anticipate some of the more common challenges peculiar to human experiments by studying criticisms of past experiments and becoming familiar with the many ways that human experiments

---

[1] Martin T. Orne, "Demand Characteristics and the Concept of Quasi-controls," in Robert Rosenthal and Ralph L. Rosnow (eds.), Artifact in Behavioral Research, Academic Press, New York, 1969.

[2] For a discussion of Hawthorne and placebo effects in various settings, see Robert Rosenthal and Lenore Jacobson, Pygmalion in the Classroom, Holt, Rinehart and Winston, New York, 1968.

can go wrong.[1] With the exception of Project 100,000,[2] few large-scale experiments have been conducted in the military to test personnel policies. Thus there are few precedents, and military leaders may be reluctant to undertake the risks that they see in using experiments to evaluate personnel programs. However, if the guidelines in the next section are followed in designing these experiments, many of the possible challenges to the experiment will have been eliminated. Also, the case studies of experiments given later in this paper provide considerable guidance on how and how not to conduct these experiments.

---

[1] For a critical review of the Head Start experiment, see Marshall S. Smith and Joan S. Bissell, "Report Analysis: The Impact of Head Start," Harvard Educational Review, Vol. 40, Winter 1970, pp. 51-104. The same issue contains a rejoinder to these criticisms by Victor G. Cicirelli, John W. Evans, and Jeffry S. Schiller. Stanley Schor indicates the vulnerability of rather carefully designed clinical trials in "The University Group Diabetes Program; A Statistician Looks at the Mortality Results," Journal of the American Medical Association, Vol. 217, No. 12, pp. 1671-1675. The same issue of JAMA contains a rejoinder by Jerome Cornfield. Many other studies are mentioned in the references cited in the two previous footnotes.

[2] Project 100,000 was an experiment to test the feasibility of relaxing mental and physical qualifications for enlisted men. For a report on this project, see Assistant Secretary of Defense (M&RA), Project One Hundred Thousand, Department of Defense, Washington, D.C., December 1969.

GUIDELINES FOR MANPOWER EXPERIMENTATION

Before an experiment is undertaken the experimenter should assure that the following conditions are met:

1. The objectives of the experiment are clearly stated. Why is the experiment being considered? How is the information going to be used? What are the major policy issues? Do the policy decisions depend upon estimates of certain key parameters? What other information is being sought?

A precise specification of the objectives of the experiment, as well as some indication of their relative importance, is essential because so many aspects of the experiment design depend on it. In some instances, the objectives of the experiment will be clear-cut and easily stated from the start, as in the case studies reported in this paper. As an example to indicate the complexities that are often involved in specifying the experimental objectives, the Navy is planning an experiment to examine the feasibility of giving all enlisted men a basic allowance for subsistence (BAS), thereby permitting them to select and pay for the food they want to eat. Presumably the policymakers in this case are Navy and DoD officials who will make recommendations to Congress, the ultimate policymaker. Is the experiment being conducted primarily to determine whether the Navy should adopt a BAS policy for enlisted men? If so, on what bases will this decision be made--relative costs, military preparedness (e.g., sick call rates, job performance measures, etc.), measures of nutritional intake under alternative food delivery schemes, or results of opinion polls to determine how morale is affected? Is the experiment being conducted to (i) determine what the BAS allowance should be, (ii) compare the enlisted men's food intake under alternative schemes, (iii) assess possible changes in health status? Clearly, these questions must be answered before the experiment can be designed. In this case, the Navy has elected to conduct a pilot study to determine enlisted men's nutritional intake before undertaking a more comprehensive experiment.[1]

---

[1]Rosemary Purcell, "Studies Presage Feeding Test," Navy Times, March 26, 1975, p. 3.

2. The treatments are prescribed in detail. The description of the treatments will include a specification of the factors that will be varied systematically, techniques and materials to be used, subpopulations to be studied, and so forth. This specification is needed to guide the choice of the experimental design and to establish the relationship between the treatments under test and the responses of interest. Given this information, the experimenters can then attempt to control other factors (background variables, environmental factors, etc.) so that (a) these other factors do not confound the main effects of the treatments, (b) information on the effects of these other factors can also be obtained.

3. The performance measures are agreed upon by the experimenters and the policymakers who will be using the information. The importance of this condition is clear from the BAS experiment above. How is a person's nutritional intake to be measured? What measure, if any, should be used to indicate how well balanced the meals are? If the primary criterion is health status, how is this to be measured?

4. An outline of the analysis that will be performed has been agreed upon. The reason for specifying this condition before the design elements are discussed is that many aspects of the experimental design (size of treatment groups, duration of the test, allocation of subjects) depend upon the type of analysis that will be carried out. This does not mean that all the details of the analysis must be spelled out in advance; the data analysts should be given some freedom to explore the experimental data for outliers, unsuspected relationships among variables of interest, and differences in responses among subpopulations that may lead to a better understanding of the process under study.

5. Treatment and control groups have been set up to assure valid estimates of the treatment effects. The purpose of the control group is to provide a standard for drawing inferences about the treatment effects. Here we are implicitly assuming that the treatment effects are to be estimated by comparing the average responses of the individuals in the treatment groups with those in the control group, perhaps after making allowances for possible inequities among the groups.[1] The appendix contains a technical discussion of some of the statistical considerations involved in designing and analyzing experiments of this type. The rationale behind the analysis rests on the premise that the differences in responses among the groups are entirely attributable to three factors: (a) the treatments themselves, (b) certain inequities among the groups, and (c) random errors. Insofar as possible, steps should be taken to minimize the differences among the groups that are attributable to (b) and (c) so that the differences that remain are primarily due to differences among the treatments.

6. The treatment and control groups are representative of the target population. Ordinarily, the experimental units will have several known characteristics (background variables, prior performance measures, etc.) that may be related to their experimental responses. In order to assure that the observed differences among the treatment groups are solely attributable to the treatments, it is important that distributions of the characteristics of the experimental units should be similar over all the treatment and control groups. The statistical rationale for balancing the groups in this sense is given in the appendix, but a more important reason for adhering to this principle is that the entire credibility of the experiment can be challenged on the basis of inequities among the groups.[2] Ordinarily, an approximate balancing of

---

[1] Not all controlled experiments are of this type. For example, the Navy may elect to use a "before-and-after" experiment to see how adopting a BAS may affect enlisted men's eating habits. Here, the same group of men is observed before and after the BAS is provided to determine changes in their nutritional intakes.

[2] See references by Smith and Bissell, Schor, and Cornfield cited in earlier footnote.

the treatment and control groups can be achieved by assigning the experimental units to the groups at random, in which case the result of the randomization should be checked for marked inequities on the known variables. If the treatment groups are small or if certain operational constraints preclude the use of randomized groups, the experimenter may deliberately assign the units to achieve balance across groups, as was done in the variable tour experiment in the Air Reserve Forces.

7. Steps have been taken to assure that the effects of the treatments will not be masked or distorted by other factors. The concern here is to guard against all factors that might inflate or reduce the experimental responses of the individuals in the treatment groups so that comparisons with the control group are invalidated. The ideal way to guard against these distortions is to guarantee that the experimental environment for the control group is identical to that for the treatment groups. Thus, in medical trials to assess the effectiveness of drugs, placebos are administered to the members of the control group that are similar in appearance to the pills containing the drugs. To make the experimental conditions even more comparable and to eliminate any effects that the experimenter might introduce, medical researchers often use "double-blind" experiments in which neither the subjects nor the experimenters know which persons receive the placebos.[1]

8. The size and/or duration of the experiment is sufficient to meet the experimental objectives. Experiments are usually conducted to estimate the treatment differences or other parameters with prespecified levels of precision. Statistical calculations of the type given in the appendix may be required to check that the sample sizes (and perhaps the duration of the test) are sufficient to meet the experimental objectives. Alternatively, the experiment may be conducted to test certain hypotheses, in which case one needs to know that the tests will have sufficient power to detect alternatives of interest. Since many aspects of the experimental design (sample sizes,

---

[1] For an interesting history of the role of the placebo in medical research, see Arthur K. Shapiro, "A Contribution to a History of the Placebo Effect," Behavioral Science, Vol. 5, No. 2, April 1960, pp. 109-135.

duration of test, choice of experimental subjects, and assignment of subjects to treatment groups) may involve considerations of a technical nature, it is advisable to have the services of a statistician who has considerable experience in experimental design.

9. Provisions for gathering reliable, comprehensive data have been made. The data should be gathered as unobtrusively as possible, preferably by disinterested observers. The control group responses should be measured in exactly the same way as the treatment group responses. Relevant background information on all participants should be gathered before the experiment gets under way. Often it is advisable to take a pretest measurement on the response variable (or something like it); this may serve as a proxy for unknown background variables that affect the process under study in a complex way.

10. The experimenters, data analysts, and all others associated with the experiment will not prejudice the findings. If possible the persons who design, monitor, and analyze the experiment should be disinterested third parties. It may be advisable to have expensive, complex experiments analyzed by two or more statisticians working independently. To facilitate reanalysis by others, provisions should be made for publishing the complete data set as part of the final report or, if the data set is too large, making it available in a machine-readable format.

11. The experiment is important enough to justify the costs. As this listing indicates, conducting a controlled experiment properly can be an expensive, time-consuming exercise involving many people other than the experimental subjects. Will the information to be gained be worth the time and effort? Should the implementation of the program be delayed until the experimental evidence becomes available? These are difficult questions, because the value of the information to be gained will ordinarily not be known until the experiment is completed.

Sometimes the costs and delays associated with experimentation are so small compared to the possible negative effects of an ill-conceived program that there is a clear obligation to experiment first. The problem in applying this principle is that the people who are most knowledgeable about the process under study may not foresee the possible negative effects, and there may be no consensus among the experts about the need for experimentation. This was the case in the experiments described in the next section.

THE VARIABLE TOUR EXPERIMENTS

With the sharp drop in nonprior service enlistments in the reserves in early 1973 following the elimination of the draft in late 1972, the services appealed to the Department of Defense to cut the term of enlistment for reservists from six years to three, contending that the six-year term of enlistment was a major impediment to recruiting. Rand researchers who had been studying the "reserve problem" questioned the wisdom of doing this, because it was not clear that shortening the term of enlistment would stimulate recruiting enough to offset the man-year losses that would result later on. We felt that there might be other recruiting incentives that would produce better results at lower cost.

Accordingly, we recommended that the Air Reserve Forces conduct a small-scale controlled experiment that would permit us to assess the attractiveness of shorter enlistment tours to potential recruits. To minimize the negative effects of adopting a shorter tour in the event that the response fell far short of expectations, we proposed that the test be conducted by permitting a few carefully selected Consolidated Base Personnel Offices (CBPOs) to offer either a three- or four-year enlistment tour for a limited period of time to see if those units would attract substantially more recruits than those recruiting using the usual six-year term. The Air Force accepted our recommendation and asked DoD for authorization to proceed with the test. DoD approved our experimental design but replaced the three-year tour by the "3x3 option" (three years of active reserve participation followed by three years in the Individual Ready Reserve) and the four-year tour by the "4x2 option" (four years of active reserve participation followed by two in the IRR).

Meanwhile Army officials, who apparently regarded the notion of waiting several months to evaluate the options as a ridiculous waste of time, were seeking authorization to reduce the enlistment term in the Army reserves to three years across the board. Undoubtedly our arguments for conducting a test in the Air Reserve Forces were important considerations behind the DoD decision to not only defer the Army proposal

but to require that the Army conduct a similar test. The guidelines that DoD specified for the Army to follow in conducting their experiment were clearly patterned after the experimental design that we had proposed for the Air Force Reserves except that the options were to be tested on a much wider scale.

The implications of adopting a shorter enlistment tour in the reserves are indicated by the following considerations:[1]

1. It is estimated that, under current retention rates in the reserves, one would need approximately 45 percent more three-year enlistees to maintain the same size steady-state force as one manned by a six-year enlistment. Hence, training costs, which are roughly proportional to the number of enlistments, would run about 45 percent higher under a three-year enlistment.

2. Since three-year enlistees would have a shorter average tour length, a smaller proportion of them would reach the higher pay grades. But pay and allowances per reserve man-year would run 10 percent higher for the three-year enlistees than for the six-year group. The reason for this apparent paradox is that personnel costs for reservists are disproportionately high during the initial period of active duty for training.

3. The experience level of the reserve forces would fall. In a steady-state force maintained entirely by nonprior service enlistments, approximately 60 percent of the men would have less than three years of service, and 21 percent would have less than one year. The corresponding percentages for the six-year group are 42 and 15.

In this case the objectives of the experiment were clear, namely, to determine whether shortening the term of enlistment would stimulate

---

[1]The figures cited below are taken from Gus W. Haggstrom, The Variable Tour Experiment in the Army Reserve Components, The Rand Corporation, R-1568-ARPA, 1975. They are based on the assumptions that both the three- and six-year groups will have an annual attrition rate of 5 percent until the initial tour is completed, a 25 percent reenlistment rate, and an annual attrition rate thereafter of 10 percent until retirement at 25 years of service. The present reenlistment rate in the Army National Guard is close to the 25 percent figure used in these calculations. Corresponding calculations for other retention rates are given in the above report.

recruiting enough to offset the disadvantages cited above. We recommended that the Air Force test both three-year and four-year enlistments, since there was little to be lost and much to be gained in testing both levels. Also, this would allow us to estimate the difference that the extra year's commitment would have upon the number of enlistments.

We recommended that the Air Force conduct the test by letting a small number of carefully selected reserve units offer the special options for six months or longer. At that time there were 125 Consolidated Base Personnel Offices (CBPOs) in the Air Reserve Forces--91 in the Air National Guard (ANG) and 34 in the Air Force Reserves (AFRES). To minimize the negative effects of the shorter enlistment options in the event that they did not prove sufficiently attractive to new recruits, we recommended that the three-year option be tested at only 10 CBPOs (five in each component) and that the four-year option be tested at 10 other CBPOs. Hence, only about one of every six CBPOs were to receive the options under our plan, with the other units serving as a control group.

The CBPOs that were permitted to offer the options were selected so that they would be representative of the entire set of CBPOs in terms of the geographical distribution of the CBPOs and the size and income level of the young male population in the vicinity of the units. Insofar as possible, we attempted to assign the options to CBPOs that were well separated from the other units to preclude the possibility that a man might enlist in a CBPO offering a shorter enlistment in lieu of enlisting for six years in a unit nearer his home. Except for some coin-flipping to choose among certain CBPOs for inclusion in the treatment groups and to decide which groups would receive the three- and four-year schemes, treatment groups were not chosen randomly in this case. The small sample sizes and the operational constraints did not permit it.

Air Force officials asked that the treatment groups be enlarged slightly to include two other ANG CBPOs and to include a sixth AFRES CBPO in the four-year group that they felt was too close to an ANG CBPO that had been selected to offer the four-year option. Otherwise, the Air Force officials endorsed our treatment group selections as being

relatively representative in terms of their strength statistics and their missions, as well as of the criteria that we had used.

The experiment began on June 1, 1973, and ran for seven months. To assure the reliability of the experimental data, we requested that the Air Reserve Personnel Center supply us computer tapes at the end of each month listing all new nonprior service (NPS) recruits by name, social security number, CBPO designation, and other information. Thus, only those recruits were counted who completed all their preinduction tests satisfactorily and were assigned to a unit.

All recruiters were informed that certain CBPOs were permitted to offer potential recruits the option of enlisting for a shorter term of enlistment. They were also informed that The Rand Corporation was conducting a study of recruiter productivity during the same period, so that all recruiting performances would be monitored closely throughout the experimental period. The recruiter supervisors were asked not to put undue pressure on the CBPOs that offered the enlistment options; to the best of our knowledge, they cooperated fully.

The experimental results indicated that the shortened enlistment options had little or no effect upon recruiting performances in the Air Reserve Forces. In comparing recruiting performances across units, we defined the "enlistment rate" for any CBPO or group of CBPOs to be the number of male NPS recruits per thousand authorized strength. (The reason for excluding female enlistments in defining the enlistment rate is that the initial tour of duty for women in the reserves was three years before the experiment began.) The eleven CBPOs that offered the 3x3 option had an overall enlistment rate of 10.2 for the seven-month period, as compared with 10.7 for the 4x2 group, and 10.5 for the 6x0 (control) group. A more detailed analysis, using analysis of covariance techniques to correct for certain inequities among the groups, indicated that the 3x3 and 4x2 groups outperformed the control group by a narrow, but statistically insignificant, margin.

Whereas most of the guidelines for conducting manpower experiments were followed in designing and conducting the variable tour experiment

in the Air Reserve Forces, the corresponding Army test that began a month later was conducted in such a way that the experiment's credibility was discredited from the start. In all fairness, the Army officials that I talked to about the test were convinced that a shorter enlistment tour was essential for manning the reserves in the absence of the draft or sizable monetary incentives. They were genuinely concerned that a delay for experimentation would contribute to a further decline in reserve strength that would have to be made up in some other way. I shared their concern, because six years must seem like an awful long time to an 18-year-old.

On the other hand, the men who join the reserves are only a small proportion of the total college-age population. The motives of these men are surely quite different from those of the majority of the population. In the absence of hard evidence to support our opinions, how could we know that the six-year term was a major factor behind the recruiting shortfalls?

Regrettably, the Army conducted their test in such a way that, even if the experimental evidence had fully supported the Army's claim, it could be challenged on many grounds. The major flaw in design was to offer the 3x3 and 4x2 options on such a wide scale. All Army reserve units in the following states were permitted to offer the shortened enlistment options:

3x3	4x2
Connecticut	Arizona
Florida	California
Hawaii	Delaware
Louisiana	District of Columbia
Massachusetts	Kansas
Mississippi	Maryland
New Jersey	Missouri
New Mexico	Nebraska
Ohio	Nevada
Oregon	North Carolina
Pennsylvania	North Dakota
Rhode Island	South Carolina
Texas	Virginia
Washington	
West Virginia	
Wisconsin	

The states not listed above and Puerto Rico served as a control group for the experiment by enlisting male recruits under the usual six-year (6x0) commitment. The Marine Corps Reserve also participated in the variable tour experiment during the second half of 1973 by offering the 3x3 and 4x2 options in the same states as the Army Reserve Components. Their experience with the options seemed to be similar to that of the Air Force in that there appeared to be no significant differences in recruiting performance among the three groups. However, there were only a few enlistments in the Marine Corps Reserve during this period, and the experimental results may have been confounded by a change in the Marine Corps recruiting program shortly after the experiment began.

By offering the 3x3 and 4x2 options on such a wide scale, the reserves were running the risk that these options might not stimulate recruiting enough to offset the man-year losses cited earlier, in which case the net effect of offering the options would be to saddle the reserves with a large group of short-term enlistees. Since many of these enlistees will be leaving the reserves at the same time that the reserves will be depleting its present supply of draft-induced volunteers, offering the options in approximately two-thirds of the states may add to their difficulties later, when the reserves will be facing even more critical manning problems.

In putting approximately a third of the states in each of the experimental groups, Army officials were only following guidelines for the experiment specified by DoD. However, they freely admitted that, subject to these guidelines, they tried to put the states that were "hurting" the most for recruits into the 3x3 and 4x2 groups. Thus, the 3x3 states were 12 percent below their authorized enlisted strengths at the start of the experiment, whereas the 4x2 and 6x0 states were only 9 and 6 percent understrength. In general, one would expect that recruiters for reserve units that are either overstrength or close to it would be under less pressure to recruit, and they would probably tend to be more selective. As an indication of how this factor may have affected recruiting in the Army National Guard (ARNG), the five 6x0 states that were more than 10 percent understrength at the start of the experiment showed a 13.8 percent increase in NPS enlistments during the

second half of 1973, whereas the six 6x0 states that were overstrength at the start showed a 46.5 percent decrease.

The wide-scale nature of the test may also have prejudiced the experimental findings. Recruiters in all states were informed at the beginning of the experiment that the Army would conduct a 90-day test of the options with approximately one-third of the states offering each of the options. Given this information, a recruiter in a 6x0 state would have good reason to expect that he would have one of the options to offer at the end of 90 days. Why not give a potential recruit a break and tell him that he can sign up later for a shorter term? Would the Army offer the options on such a wide scale if there were any chance of not reducing the tour of duty for all new reservists? The anticipated flow of mail to Congress when the 6x0 recruits went on active duty for training and learned that most of their buddies from other states had enlisted for shorter tours of duty would surely be sufficient to force the military to offer all recruits the same option.

Given the publicity associated with the experiment and the factors already mentioned, we questioned whether the 6x0 states could really be treated as a control group in the usual sense. Efforts should have been made to guarantee that the 6x0 recruiters were performing up to par during the experimental period. At the very least, the recruiters should have been informed that their performances would be monitored more closely during the experimental period. A comparison of the recruiting performances of the 6x0 states during the experimental period with their performances both before and after the experiment indicate that the 6x0 states were not performing up to par during the experiment. Perhaps the recruiters themselves were blameless. There were unverified reports that some enlistees were crossing state lines to join units offering the 3x3 option with the hope of transferring back after active duty for training.

To confound the analysis further, the ARNG conducted some marvelously productive recruiting campaigns in certain states during the experimental period. Of course, the ARNG elected to concentrate most of its efforts in 3x3 states. Many ARNG technicians who were not

ordinarily engaged in recruiting activity took time off from their other duties to work as recruiters. The most successful campaigns were conducted in the four 3x3 states listed below. The recruiting performances during the months of the campaigns are marked with asterisks.

	Male NPS Enlistments						
State	July	Aug.	Sep.	Oct.	Nov.	Dec.	Total
Louisiana	106	63	67	79	347*	86	748
Massachusetts	24	48	33	263*	80	37	485
New Jersey	65	128	127	338*	68	34	760
Wisconsin	8	11	12	5	9	286*	331

Wisconsin's December performance of 286 recruits is particularly noteworthy, given that the state averaged only nine recruits per month during the other five months.

Of the 2324 male NPS enlistments in these four states during the experimental period, over half were obtained during the single months in which the recruiting campaigns were conducted. Moreover, these four states accounted for almost one-half of the 4855 enlistments in the 3x3 states during the six-month period. Intensive campaigns were also conducted in a few 4x2 and 6x0 states, but the 3x3 group benefited far more from this activity.

By conducting these intensive campaigns primarily in the 3x3 states, the Guard effectively destroyed the credibility of the experiment insofar as establishing the worth of the enlistment options. If the analysis had shown a 200 percent increase in enlistments due to the 3x3 options, no one would have believed it because of this factor. Presumably, the 3x3 states were chosen for most of the campaigns to take advantage of the option and bring more recruits into the fold. But why was Louisiana, an overstrength state, chosen for a campaign late in the experimental period when Louisiana's recruiting had been strong throughout the period? Was this a deliberate attempt to make the 3x3 option look good?

Data on the number of man-days spent on these campaigns were not available. The Army's provision of a data base for analyzing the

experiment left a lot to be desired in other respects. Although the ARNG is geared to providing personnel data on a state-by-state basis, the U.S. Army Reserve (USAR) is not. Because of the lack of a suitable system for getting enlistment data by state on a timely basis, the Army resorted to asking for "flash reports" from the individual states at the end of each month. Not only did this institute a new time-consuming report, it probably also led to clerical errors and incorrect counts. In contrast, we started receiving monthly listings of new NPS recruits into the Air Reserve Components on magnetic tape before the Air Force experiment got under way. The USAR successfully resisted our efforts to get state-by-state monthly figures on NPS enlistments for the six-month period before the experiment began. Also, they provided no information whatsoever on the amount of recruiting activity that goes on routinely in each state.

As was pointed out previously, the experimental states had substantially larger strength deficits. There were other imbalances among the treatment groups. In terms of demographic characteristics, the experimental states tended to be more populous and have higher incomes and educational attainment than the 6x0 states. The analysis of the experiment revealed that enlistment rates seem to be sensitive to the level of unemployment. Since the 4x2 states had lower unemployment rates on average during that period, the recruiters in those states may have been operating under a handicap.

In theory, one can try to make allowances for imbalances among the groups on these variables using analysis of covariance or multiple regression. (See the appendix for details.) However, if there are imbalances among the groups on key variables, the estimates of the treatment effects become much more sensitive to the way that the analytical model is specified, and even if the model is specified perfectly, the efficiencies of the parameter estimates are reduced. In other words, the analysis becomes less precise, more sensitive to anomalies in the data, and more vulnerable to the preconceptions and whims of the data analyst.

Despite all the defects of the Army experiment, it still yielded valuable information about the effects of the options. In terms of the overall enlistment rates (male NPS enlistments per thousand authorized strength) computed from the raw data, the 3x3 and 4x2 groups outperformed the 6x0 group by 60 and 47 percent respectively. Since these percentages surely exaggerate the effects of the options because of the many factors that tended to inflate the treatment effects, they provide useful upper bounds on the effects of the options.

Various statistical techniques were used in an attempt to make allowances for the inequities among the treatment groups in estimating the treatment effects. Although the effects clearly cannot be estimated with precision, it appears from this more detailed analysis that the 3x3 option resulted in a 20-40 percent increase in NPS enlistments during the experimental period, and the 4x2 option yielded a 10-30 percent increase.[1] These results indicate that adopting the 3x3 option would not attract enough recruits to offset the later man-year losses, let alone compensate for the other disadvantages of the 3x3 scheme. The estimated response to the 4x2 option in the Army reserves appears to be close to that required to offset the later man-year losses under current reenlistment rates. The 4x2 option seems preferable to the 3x3 scheme in other respects, but personnel costs would rise under both schemes, and other factors should be considered before implementing either option.

The Rand Corporation provided the military monthly progress reports during the course of the experiment. When the experimental results indicated that the Army proposal to adopt a 3x3 option across the board would probably be detrimental to the reserves, DoD stopped the experiment on December 31, 1973. A few months later, the Secretary of Defense authorized the reserves to enlist a limited number of men under 3x3 and 4x2 schemes, but this option was restricted to not more than 20 percent of the total NPS enlistments and to applicants in the higher mental categories.[2]

---

[1]The analysis supporting these estimates, as well as a complete listing of the experimental data, are given in the report cited earlier.

[2]"Shorter Hitch OKed in Selected Reserve," Air Force Times, April 17, 1974, p. 21.

As an indication of the savings achieved through deferring the implementation of the 3x3 scheme for a single year, in 1973 the Army and Air Force reserve components had approximately 25,000 male NPS enlistees of whom approximately 20,000 enlisted for the full six years. If the 3x3 scheme had been enacted across the board at the beginning of the year and if the 3x3 scheme had yielded 30 percent more enlistees (a liberal estimate), then instead of having 20,000 enlistees with a six-year commitment they would have had 26,000 3x3 enlistees. Based upon current attrition rates in the ARNG, we estimate that these 20,000 six-year enlistees will average 6.7 years of service in the active reserves for a total of 134,000 man years, whereas the 26,000 3x3 enlistees would average only 4.6 years of service for a total of 120,000 man years. Thus, in 1973 alone, by experimenting instead of adopting the 3x3 scheme across the board, the Army and Air Force reserve components will have gained 14,000 man years. This net gain of 14,000 will occur despite the fact that the reserves will forego approximately 16,000 man years over the next three years from the 6,000 additional 3x3 recruits; the reserves will compensate for this short-term loss of 16,000 man years between 1973 and 1976 by a gain of approximately 30,000 man years during the following three years, at a time when the reserves will be depleting their present supply of draft-induced volunteers.

The experiment destroyed the myth that the six-year term of enlistment in the reserves was the major reason for their recruiting difficulties. Some Army officials seemed to think that enlistments would double or triple under a three-year term. Air Force recruiters who had the 3x3 and 4x2 options to offer expressed disappointment over the lack of interest in these options on the part of potential recruits. An interesting feature of the experiment was that male recruits in the 3x3 and 4x2 states could enlist for the full six-year term if they wished. To our surprise, approximately one-third of the Army recruits in the 3x3 states enlisted for a full six years, and almost two-thirds of the Air Force enlistees in the units offering the 3x3 options signed up for six years. Also, the pattern of enlistments across states and units from month to month indicated that other factors

(recruiting campaigns, unemployment rates, unusually effective recruiters) play a more dominant role in the recruiting process.

The failure of the 3x3 scheme to live up to its expectations will lead the military to consider other recruiting strategies that might have been overlooked had the experiment not taken place. Given the experimental results, the services may want to take another look at the 4x2 scheme, or they may consider a variable enlistment bonus that pays long-term enlistees more than the short termers. Our analysis showed that the Army conducted some amazingly productive recruiting campaigns in certain states during the experimental period. They may want to investigate whether such campaigns are more cost-effective in the long run than implementing a shorter term of enlistment.

The experiment, despite its flaws, set a precedent for using controlled field studies to evaluate enlistment incentives. There are many lessons to be learned from the experiment, but perhaps the main lesson is that a lot of valuable information about the recruiting process and the attractiveness of incentives can be gained from an experiment of this type. It is our hope that the Air Force test will serve as a prototype for similar experiments in the future and that the lessons learned from the Army experiment will not be wasted.

APPENDIX

SOME ASPECTS OF EXPERIMENTAL DESIGN IN ANALYSIS OF COVARIANCE SITUATIONS

This appendix provides a discussion of some of the technical details associated with designing controlled experiments to be analyzed using analysis of covariance models. It is intended primarily for readers having a good theoretical background in the field of linear models or analysis of variance.[1]

Suppose an experiment is to be conducted with p treatment groups--including a control group, if there is one--for the purpose of estimating the treatment effects and certain other parameters. To specify desiderata for experimental designs, we begin by defining our terms, stating our assumptions, and considering how an already completed experiment of this type might be analyzed.

Suppose there are n participants altogether and that the number of individuals assigned to the p groups are $n_1, n_2, \ldots, n_p$. Let y_i be the response of the ith participant, and let $z_i = (z_{i1}, z_{i2}, \ldots, z_{ih})'$ be a vector of characteristics for this participant such that, in the absence of the treatments, the expected value of y_i is an unknown function of z_i, say

$$\eta_i = E(y_i) = f(z_i).$$

---

[1] A definitive reference in this field is Henry Scheffè's The Analysis of Variance, John Wiley & Sons, New York, 1959. See especially Chapter 6, The Analysis of Covariance.

Thus, in the absence of the treatments, the responses can be written in the form

$$y_i = \eta_i + e_i = f(z_i) + e_i$$

where e_i is the random error associated with y_i.

Now consider the response y_i of the ith individual if he is put in the jth treatment group. If his response measured without error is μ_{ij} then the effect of the jth treatment for that individual is defined to be $\tau_{ij} = \mu_{ij} - \eta_i$. Again letting e_i denote the random error associated with y_i, we have that

$$(1) \qquad y_i = \Sigma_{j=1}^{p} \tau_{ij} x_{ij} + f(z_i) + e_i,$$

where $x_{ij} = 1$ if the ith individual is assigned to the jth group and 0 otherwise. It will be assumed below that the errors e_i are uncorrelated random variables with mean 0 and variance σ^2.

For the moment we shall impose two further assumptions which are somewhat restrictive in nature. Having deduced the design implications when these assumptions hold, we shall then return to the more general formulation above.

Assumption A. The jth treatment yields the same additive effect τ_{ij} for all individuals, i.e., the subscript i on τ_{ij} can be omitted. In this case, the parameter τ_j is called the jth treatment effect.[1]

Assumption B. The function $f(z_i)$ is linear in the components of z_i, i.e.,

$$f(z_i) = \alpha + \gamma' z_i$$

[1] If the values of τ_{ij} of the jth treatment vary from individual to individual, the jth treatment effect will be defined as the average of the individual effects over all individuals in the target population.

where α and $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_h)'$ are unknown parameters.

Under the above assumptions, equation (1) can be written as a linear model

$$y_i = \Sigma_{j=1}^{p} \beta_j x_{ij} + \gamma' z_i + e_i \tag{2}$$

where $\beta_j = \tau_j + \alpha$. Assuming that the resulting design matrix is of full rank, it follows from the Gauss-Markov Theorem that the best linear unbiased estimators (BLUEs) of the parameters β_j and γ_k in this model are the ordinary least-squares regression coefficients $\hat{\beta}_j$ and $\hat{\gamma}_k$.

Several comments are in order:

1. There are no restrictions on the components of z_i except that the resulting linear model have full rank. They may be personal characteristics, environmental variables, block indicators, transformations of these variables, etc. For the moment, we assume they are known constants.

2. If the pth treatment group is a control group, then $\tau_p = 0$ and the BLUEs of the other treatment effects are $\hat{\tau}_j = \hat{\beta}_j - \hat{\beta}_p$. More generally, if $\psi = \Sigma c_j \tau_j = c'\tau$ is any <u>contrast</u> in the treatment effects (i.e., $\Sigma c_j = 0$), then $\psi = c'\beta$, and the BLUE of ψ is $\hat{\psi} = c'\hat{\beta}$.

3. The treatments may correspond to different levels of two or more factors in a two-, three-, or higher-way layout. In this case the "main effects" and "interactions" are again contrasts in the β_j's.

It is often asserted that, in comparing treatments using analysis of covariance, the individuals should be assigned to the treatment groups so that the covariate values z_{ik} are "balanced" across groups. Let $\bar{z}_j = (\bar{z}_{j1}, \bar{z}_{j2}, \ldots, \bar{z}_{jh})'$ be the mean vector of the covariate values within the jth group, i.e., $\bar{z}_{jk} = \Sigma_{i=1}^{n} z_{ik} x_{ij}/n_j$ for $k = 1,2,\ldots,h$.

One way to specify balance is to require that the covariate mean vectors be the same in all treatment groups, i.e.,

$$\bar{z}_1 = \bar{z}_2 = \ldots = \bar{z}_p \quad . \tag{3}$$

Ordinarily, the covariate values are such that perfect equality cannot be achieved, in which case approximate equality is prescribed. Why? The reason usually given is to reduce bias, a concern that will be examined later.

If Assumptions A and B leading to the linear model (2) are satisfied, the least-squares estimators of the contrasts in the β_j's are unbiased no matter how unbalanced the treatment groups are. Hence, bias is not the key issue here provided that these assumptions are satisfied. But, even if these assumptions are fully satisfied, there are still good reasons for prescribing balance in the sense of (3) above.

Theorem. Consider a linear model specified by (2) with fixed treatment group sizes $n_1, n_2, \ldots, n_p$. Any experimental design in which the experimental units are assigned to the treatment groups in such a way that (3) holds is optimal in the sense that, if $\hat{\beta}$ and $\hat{\gamma}$ denote the least-squares estimators of β and γ,

(a) the variances of all estimated contrasts $\hat{\psi} = \Sigma c_j \hat{\beta}_j$ are simultaneously minimized;

(b) the variances of the regression coefficients $\hat{\gamma}_k$ are minimized.

Proof: The proof depends on standard facts that can be found in Scheffè's text.[1] The least-squares estimator of β_j is

$$\hat{\beta}_j = \bar{y}_j - \hat{\gamma}'\bar{z}_j. \tag{4}$$

---

[1] Scheffè, ibid, Chapter 6. Scheffè gives equations of the same type as (6) for the case of one or two covariates on pp. 209 and 213.

Hence, for any contrast $\psi = c'\beta$,

$$\hat{\psi} = c'\hat{\beta} = \Sigma c_j \bar{y}_j - \hat{\gamma}'a \tag{5}$$

where $a = \Sigma c_j \bar{z}_j$. Since the sample means $\bar{y}_j$ are uncorrelated with the coordinates of $\hat{\gamma}$,

$$\mathrm{Var}(\hat{\psi}) = \sigma^2(\Sigma c_j^2 n_j^{-1}) + \mathrm{Var}(a'\hat{\gamma}). \tag{6}$$

It follows that $\mathrm{Var}(\hat{\psi})$ is minimized by assigning the experimental units in such a way that $\mathrm{Var}(a'\hat{\gamma})$ is as small as possible. Since $a = 0$ under (3), part (a) of the theorem is proved.

To prove part (b), we first reparameterize the model (2) to

$$y_i = \Sigma_{j=1}^{p} \alpha_j x_{ij} + \gamma' w_i + e_i$$

where $\alpha_j = \beta_j + \gamma'\bar{z}$, $\bar{z} = \Sigma z_i/n$, and $w_i = z_i - \bar{z}$. Partitioning the resulting design matrix into $X = [U\ W]$, we see that the condition (3) is equivalent to requiring that the columns of U be orthogonal to the columns to W. Let W_k denote the kth column of W, and let P denote the projection operator on the other columns of X. Then $\mathrm{Var}(\hat{\gamma}_k) = \sigma^2/||QW_k||^2$ where $Q = I - P$. Since

$$||QW_k||^2 = ||W_k||^2 - ||PW_k||^2 = ||W_k||^2 - ||P_1W_k||^2 - ||P_2W_k||^2$$

where P_1W_k is the projection of W_k upon the other W_j's and P_2W_k is the projection of W_k upon the orthocomplement of the column space of W in the column space of X, minimizing $\mathrm{Var}(\hat{\gamma}_k)$ amounts to choosing the columns of U to minimize $||P_2W_k||^2$. Under (3), P_2W_k becomes the projection of W_k onto the column space of U, which is 0 since W_k is orthogonal to the columns of U. Thus, (3) minimizes $\mathrm{Var}(\hat{\gamma}_k)$ for each k.

To paraphrase the theorem, if the linear model assumptions hold, the most precise (least variable) linear unbiased estimators of the treatment differences and other contrasts are obtained by balancing the treatment groups in the sense of (3), in which case the regression coefficients γ_k are also estimated with maximum precision.

Note from the proof of the theorem that the BLUE of any contrast $\psi = c'\beta$ (or $c'\tau$) is $\hat{\psi} = c'\hat{\beta}$ which, by equation (5) reduces to $\hat{\psi} = \Sigma c_j \bar{y}_j$ when (3) holds. Thus, although the estimates $\hat{\beta}_j$ depend on the covariate vectors z_i, the estimates of the contrasts in the β_j's (or τ_j's) depend only on the group means, not on the z_i's, when the groups are balanced. The estimator $\hat{\psi}$ is exactly the same estimator that would be used if the z_i's were omitted in the specification of (2). In addition, as long as (3) is satisfied, the estimates of the treatment differences are unaffected by: (i) omitting some of the covariates in the analysis, perhaps unknowingly; or (ii) including too many irrelevant covariates.[1]

The estimates of the treatment effects may still be biased by the omission of an important covariate or by including the "wrong" transformation of a covariate. Ideally, the solution is to have the treatment groups balanced in the sense of (3), not only for the h covariates in the model, but for transformations of these covariates and for

---

[1] These arguments should not be construed as favoring the omission of the covariates when (3) holds. The regression coefficients γ_k will probably have independent interest. Also, the covariates need to be included to estimate σ^2 and to determine the standard errors of estimates.

other variables that might have been included. One method of achieving such a balance, at least approximately, when the treatment groups are large is to assign the individuals to the groups either completely at random or randomly within blocks (or strata) determined by partitioning the subjects according to certain covariate values. Checks can then be performed on the known covariates to verify that the covariate means and variances are approximately equal across groups. If the treatment groups are not approximately balanced, a second randomization can be performed. Alternatively, if the groups are small, one can deliberately assign the individuals to the groups in such a way as to achieve balance, but there are dangers here because the experimenter's choices may unconsciously favor some treatment.

Since the assumptions leading to the linear model (2) are not likely to be satisfied perfectly, it is of interest to determine how well the estimators obtained using (2) might perform when the true model should be the more general one specified in (1), namely,

$$y_i = \Sigma_{j=1}^{p} \tau_{ij} x_{ij} + \eta_i + e_i$$

where $\eta_i = t(z_i)$ and the errors e_i are uncorrelated with mean 0 and variance σ^2.

Consider estimating the treatment difference $\psi = \beta_1 - \beta_2$ (or, equivalently, $\tau_1 - \tau_2$) using the BLUE from model (2):

$$\hat{\psi} = \hat{\beta}_1 - \hat{\beta}_2 = \bar{y}_1 - \bar{y}_2 - \hat{\gamma}'(\bar{z}_1 - \bar{z}_2). \tag{7}$$

The estimated value of this estimator in model (1) is

$$E(\hat{\psi}) = (\bar{\tau}_1 - \bar{\tau}_2) + (\bar{\eta}_1 - \bar{\eta}_2) + \delta'(\bar{z}_1 - \bar{z}_2), \tag{8}$$

where $\bar{\tau}_j = \Sigma_{i=1}^{n} \tau_{ij} x_{ij}/n_j$, $\bar{\eta}_j = \Sigma_{i=1}^{n} \eta_i x_{ij}/n_j$, and $\delta = E(\hat{\gamma})$.

This calculation of (8) was carried out under the assumption that the values τ_{ik} and η_i are fixed constants. If the individuals are assigned to the treatment groups at random, the quantities $\bar{\tau}_j$, $\bar{\eta}_j$, and $\bar{z}_j$ can be considered as random variables having means $\tau_j = \Sigma_{i=1}^{n} \tau_{ij}/n$, $\eta = \Sigma\eta_i/n$, and $\bar{z} = \Sigma z_i$ respectively. Thus, the expected value of the right member of (8) is $\tau_1 - \tau_2$. In this sense, applying model (2) to randomly chosen treatment groups yields unbiased estimates of the average treatment differences. However, few statisticians would take this argument very seriously because, once the randomization of subjects to treatments has been accomplished, the beautiful symmetry and balance of the random assignment process are replaced by the homely inequities of the actual assignment. The only outcome of the random assignment that matters is the one that occurred; proving theorems by averaging over all the random assignments that did not occur constitutes a meaningless exercise.

Returning to equation (8), we can ask how the z_i's should be assigned to the two groups to assure that $E(\hat{\psi})$ is approximately equal to $\tau_1 - \tau_2$, where τ_1 and τ_2 are the average treatment effects defined as in the previous paragraph. It is sufficient to require that each of the groups be "representative" of the target population so that (i) $\bar{\tau}_1$ and $\bar{\tau}_2$ are approximately equal to τ_1 and τ_2, (ii) $\bar{\eta}_1$ and $\bar{\eta}_2$ are approximately equal, and (iii) $\bar{z}_1$ and $\bar{z}_2$ are approximately equal, in which case the third term in (8) can be dropped.

It is hard to define what the words "approximately equal" in these conditions mean, but the implication is that, insofar as possible, the distributions of the z_i's in the two treatment groups should be

approximately the same as that in the target population. Since the individual treatment effects τ_{ij}, as well as the values η_i, can be assumed to be functions of the individual's characteristics z_i, assuring that the z_i's have approximately the same distribution in each of the treatment groups as in the target population will go a long way toward eliminating any bias in (8). Completely random or stratified random samples can be used to achieve this if the treatments are large, but the results of the randomization should be checked. A reasonable check of this condition is provided by comparing means, variances, and correlation coefficients of the covariate values across the treatment groups.

If the treatment groups are unbalanced on a particular covariate (say, ability level) and this variable is closely related to the response variable (say, test performance), then the term $\bar{\eta}_1 - \bar{\eta}_2$ in (8) may not be negligible relative to $\bar{\tau}_1 - \bar{\tau}_2$. However, the inclusion of the covariate values in the analysis will help offset the term $\bar{\eta}_1 - \bar{\eta}_2$, provided that $\eta_i = f(z_i)$ is approximately linear in z_i, because the second and third terms in the right member of (8) will tend to offset one another. The main concern in making deductions using treatment groups having sizable imbalances is that the individual treatment effects τ_{ij} may be quite different, say, for the low ability participants than for rest of the population. In terms of the symbolism introduced above, the concern is that $\bar{\tau}_1 - \bar{\tau}_2$ may not be close to $\tau_1 - \tau_2$. On the other hand, if

$$\tau_{ij} = \tau_j + g(z_i)$$

where $g(z_i)$ is linear in z_i, then it is easily seen that applying

analysis of covariance will again yield unbiased estimates of the treatment differences, provided that $\eta_i = f(z_i)$ is also linear in the covariates.

Thus, even if the assumptions of model (2) are not satisfied, using analysis of covariance may lead to unbiased estimates of the treatment differences, whether there are imbalances among the treatment groups in the covariate values or not. However, there are many "if's" and "maybe's" associated with this conclusion that disappear when the treatment groups are balanced.

The discussion above was predicated on the assumption that the treatment group sizes $n_1, n_2, \ldots, n_p$ are fixed beforehand, but ordinarily the experimenter will have some freedom to choose these values in designing the experiment. Of course, the idea is to specify the group sizes to enhance the precision of the analysis to the greatest extent possible subject to the operational and budget constraints that are imposed.

Ordinarily, there are certain parameters of special interest, and the experimenter will want to choose the group sizes to estimate these parameters as precisely as possible. If these parameters can be estimated without bias, the problem can often be posed as one of minimizing a weighted sum of the variances of the parameter estimators, subject to the operational and budgetary constraints.[1] Variance calculations of the type given in equation (6) become relevant in such situations. Alternatively,

---

[1]For a theoretical treatment of this topic, see John Conlisk and Harold Watts, "A Model for Optimizing Experimental Designs for Estimating Response Surfaces," in Harold Watts et al., Field Experimentation in Income Maintenance, Reprint 54, Institute of Research on Poverty, University of Wisconsin, 1970.

the experiment may be conducted to carry out certain tests of hypotheses, in which case power calculations of the type treated in Scheffè's Analysis of Variance become the key consideration.

Often the optimal group sizes will depend on values of unknown parameters for which approximate estimates are not available. For this and other reasons, one may want to consider doing a preliminary pilot study or using sequential experimentation. Some of the considerations involved are treated in books on experimental design,[1] but the peculiarities of each controlled experiment are ordinarily such that general theory provides only partial guidance.

---

[1] D. R. Cox, Planning of Experiments, John Wiley & Sons, New York, 1958; William G. Cochran and Gertrude M. Cox, Experimental Designs, Second Edition, John Wiley & Sons, New York, 1957.