OORR 58-1

Office of Ordnance Research

PROCEEDINGS OF THE SECOND CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH DEVELOPMENT AND TESTING

OFFICE OF ORDNANCE RESEARCH, U. S. ARMY
BOX CM, DUKE STATION
DURHAM, NORTH CAROLINA

OORR 58-1

Office of Ordnance Research

PROCEEDINGS OF THE SECOND CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH DEVELOPMENT AND TESTING

OFFICE OF ORDNANCE RESEARCH, U. S. ARMY
BOX CM, DUKE STATION
DURHAM, NORTH CAROLINA

REPRODUCTION QUALITY NOTICE

This document is the best quality available. The copy furnished to DTIC contained pages that may have the following quality problems:

- Pages smaller or larger than normal.
- Pages with background color or light colored printing.
- Pages with small type or poor printing; and or
- Pages with continuous tone material or color photographs.

Due to various output media available these conditions may or may not cause poor legibility in the microfiche or hardcopy output you receive.

☑ If this block is checked, the copy furnished to DTIC contained pages with color printing, that when reproduced in Black and White, may change detail of the original copy.

OFFICE OF ORDNANCE RESEARCH
Report No. 58-1
February 1958

PROCEEDINGS OF THE SECOND CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT AND TESTING

Sponsored by the Army Mathematics Steering Committee
conducted at
Diamond Ordnance Fuze Laboratories
and
National Bureau of Standards
17-19 October 1956

Office of Ordnance Research
Ordnance Corps, U. S. Army
Box CM, Duke Station
Durham, North Carolina

TABLE OF CONTENTS

* This paper was presented at the Conference. It is not published in these Proceedings.

TABLE OF CONTENTS (Cont'd)

---

* This paper can be found in a classified security information (Confidential) appendix of this Technical Manual.

** This paper was presented at the Conference. It is not published in these Proceedings.

TABLE OF CONTENTS (Cont'd)

---

* This paper was presented at the Conference. It is not published in these Proceedings.

Initial Distribution

The initial distribution list of the Proceedings of the Second Conference on the Design of Experiments in Army Research, Development and Testing includes those who attended the meeting and/or the government installations with which they are associated. For economy, only a limited number of copies have been sent to each. Additional copies will be transmitted upon request.

The First Conference on the Design of Experiments in Army Research, Development and Testing was held on October 19-21, 1955 at the Diamond Ordnance Fuze Laboratories and the National Bureau of Standards and its Proceedings have been published. On the basis of the success of this Conference the Army Mathematics Steering Committee of the Research and Development Office of the Department of the Army decided that a similar Conference should be organized and held during the fall of 1956.

Accordingly, the Second Conference was held on October 17-19, 1956 at the Diamond Fuze Laboratories and the National Bureau of Standards. The organization of the Second Conference was similar to that of the First Conference. There were three categories of sessions. The first category consisted of invited papers by well-known authorities in the design of experiments. The second consisted of technical papers contributed by research workers from the various Army research, development and testing facilities. The third category was composed of clinical sessions devoted to presentation and discussion of partially solved or unsolved problems which had arisen in these facilities. The program of the three-day conference appears on the next few pages of these Proceedings.

The Second Conference was attended by 181 registrants and participants from 67 organizations. Speakers and other participants came from the Bell Telephone Laboratories, General Electric Company, National Bureau of Standards, National Institute of Health, Princeton University, University of North Carolina, Virginia Polytechnic Institute, and 17 Army facilities.

The present volume of Proceedings contains 26 papers and an appendix which contains 3 classified papers, all of which were presented at the Conference. The papers are being made available in this form as a contribution to wider dissemination and use of modern statistical principles of the design of experiments in research, development, and testing work of concern to the Army.

The members of the Army Mathematics Steering Committee take this opportunity to express their thanks to those research workers in the various Army research, development, and testing facilities who participated in the Conference; to Lt. Colonel J. A. Ulrich, the Commanding Officer of the Diamond Ordnance Fuze Laboratories and to Dr. A. V. Astin, the Director of the National Bureau of Standards, for making available the excellent facilities of their two organizations for the Conference; to Mr. John A. Wheeler who handled the details of the local arrangements for the Conference at both installations; and to Dr. F. G. Dressel of the Office of Ordnance Research who carried through the details, including all correspondence involved in organizing the Conference and in preparing these Proceedings.

S. S. Wilks
Professor of Mathematics
Princeton University

SECOND CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH DEVELOPMENT AND TESTING
17-19 October 1956
Diamond Ordnance Fuze Laboratories
and
National Bureau of Standards

17 October 1956

On Wednesday all sessions of the Conference will take place in the East Building Conference Room of the National Bureau of Standards.

REGISTRATION: 0900 - 0930 (Eastern Standard Time)

MORNING SESSION: 0930 - 1215

Chairman: Professor S. S. Wilks
Princeton University

Introductory Remarks: Dr. Robert D. Huntoon,
Associate Director for Physics,
National Bureau of Standards

The Planning of Experiments in the Presence of Variation
Professor George E. Nicholson, Jr., University of North Carolina

The Predesign Phase of Large Sample Experiments
Dr. Carl A. Bennett, General Electric Company

LUNCH: 1215 - 1345

AFTERNOON SESSION: 1345 - 1615

Chairman: Colonel G. F. Leist, Ordnance Corps
Commanding Officer of the Office of Ordnance Research

Recent Research on Statistical Problems in Subjective Testing
Professor Ralph A. Bradley, Virginia Polytechnic Institute

Applications of Order Statistics in Medical Experiments
Drs. Bernard G. Greenberg and A. E. Sarhan,
The University of North Carolina

MIXER: 1730 (South Room, Shoreham Hotel)

CLINICAL SESSION A (Cont):

Determining Whether a Product Satisfies Taste Specifications
N. J. Gutman, Quartermaster Food and Container Institute

Experimental Designs for War Gaming
P. R. Newcomb, Operations Research Office

CLINICAL SESSION B: 0900 - 1130 - Chemistry Bldg. Lecture Room

Chairman: F. E. Grubbs, Ballistics Research Laboratories

Panel Members: Churchill Eisenhart, National Bureau of Standards
M. B. Wilk, Princeton University

Experimental Design for Determining Specification Limits for Manganese-Aluminum Bronze
S. L. Eisler, Rock Island Arsenal

Increasing the Sample Size by Rebuilding Test Material
J. W. Coy, White Sands Proving Ground

Sampling Plan for Packaging Materials Produced by a Continuous Process
S. L. Eisler, Rock Island Arsenal

Some Observations on the Use of Models in the Design of an Experiment
J. W. Mitchell, Frankford Arsenal

CLINICAL SESSION C: 0900 - 1130 - Manse Bldg. Lecture Room

Chairman: D. M. Meals, Combat Operations Research Group

Panel Members: J. M. Cameron, National Bureau of Standards
S. S. Wilks, Princeton University

Short Range Radio Scatter Propagation Survey
R. E. Lacy, C. E. Sharp, and H. G. Linder, Signal Corps Engineering Laboratories

Experimental Designs for Organization Research Using Limited Resources
R. H. Burros, Combat Operations Research Group

Some Suggested Field Experiments
Lt. Colonel W. L. Clement, Operations Research Office

Evaluation of Interlaboratory Tests with Limited Controls and Data
W. K. Murray, Watertown Arsenal

CLINICAL SESSION D: 0900 - 1130 - Avenue Annex Conference Room

Security Classification - The paper by G. J. Blakemore and W. C. Pettijohn carries a classification of SECRET, and the paper by Fred Leone is classified CONFIDENTIAL.

Chairman: J. O. Harrison, Jr., Operations Research Office

Panel Members: G. E. Nicholson, Jr., University of North Carolina
John Tukey, Princeton University and Bell Telephone Laboratories

Testing Squad Indirect Fire Weapons
G. J. Blakemore, Jr., and W. C. Pettijohn, Operations Research Office

Design of Experiment in the Development of Ballistic Systems
F. C. Leone and H. Kahn, Frankford Arsenal

Design of Experiment Procedures in Ordnance Research
A. Bulfinch, Picatinny Arsenal

AFTERNOON SESSION: 1300 - 1530 - East Building Conference Room, National Bureau of Standards

Chairman: Dr. W. J. Youden, National Bureau of Standards

Derived Linear Models in the Analysis of Variance
Dr. M. B. Wilk, Princeton University

Choice of Error in the Design of Experiments
Dr. Jerome Cornfield, National Institute of Health

AN EXAMPLE OF DESIGN OF EXPERIMENTS AT THE NATIONAL BUREAU OF STANDARDS

R. D. Huntoon
The National Bureau of Standards

I wish to extend a dual welcome to the members of the Second Joint Conference on the Design of Experiments in Army Research, Development and Testing. You are hereby welcomed to the laboratories of the National Bureau of Standards and to the Diamond Ordnance Fuze Laboratories. We both wish you every success in this your second conference.

To some of you, it may seem a little confusing that you came to DOFL for the conference and find your meeting starting off in NBS. It may help if I explain that DOFL was, until 1953, a part of NBS. At that time, the ordnance activities of NBS were transferred to the Department of Defense and DOFL was established as a facility of the Office of the Chief of Ordnance, Department of the Army. This was in some respects merely a change of title, since essentially the same people are doing the same work in the same laboratories, and we still work closely and harmoniously together as we did earlier.

The reason for this separation is interesting and worth discussing briefly, for it gives an insight into the aims and missions of the two institutions. The statutory functions of NBS, as authorized by the Congress, are six in number. They fall into two groups which I like to call direct and indirect. Stated briefly, the direct functions are:

1. Development and custody of the national standards and their dissemination via calibrations.

2. Determination of physical constants and critical properties of materials.

3. Development of methods of testing materials, mechanisms and structures.

An institution which is properly staffed and equipped to fulfill these functions in all the fields of the physical sciences is in a unique position in the Government to perform additional functions which derive from these three. The derived functions are:

4. Cooperation with other government agencies and private organizations in the development of codes and specifications.

5. Scientific and technical advice and consultation service to other government agencies.

6. Invention and development of devices to serve the special needs of the government.

Before proceeding with the discussion, it is appropriate to pause here and emphasize the fact, which should be clear from the statement of the functions, that NBS is not a consumer testing organization as is sometimes mistakenly believed. It is an institution devoted to the science of measurement as a service to the country's scientists and engineers.

The advent of the last war naturally brought great emphasis on the third number of the trilogy of derived functions, i.e., the invention and development of devices to serve the special needs of the government. During the war and the years immediately following, there grew up within NBS an institution within an institution whose mission was to perform research and development leading to end item hardware for military use. In fact, this institution already known as the Diamond Ordnance Laboratories had grown to the point where its program was larger than that of the rest of NBS. A careful study of the situation in 1953 led to the recommendation that the "Diamond Laboratories" should become a separate institution, and the recommendation was implemented. We now work together compatibly, each toward its own objectives with mutual assistance and sharing of facilities.

The importance of design of experiment is well recognized in both institutions and in fact we consult and collaborate from time to time in the design of experiment in the full technical sense of the term. We are, therefore, pleased to have this conference assemble here for we feel that our staffs will benefit from the stimulating new information and points of view which should emerge from these meetings.

And now it is interesting to turn for a few moments from the general to the specific and take a brief look at an example of design of experiment in progress in the physical constants work at NBS.

We, along with the other national standardizing laboratories of the world, are engaged in devising new experiments for a precise determination of the acceleration of gravity, g. Strictly speaking, g is not a physical constant, although it is commonly referred to as one. It varies from place to place over the surface of the earth and very slightly from time to time at any one place. However, it is essentially a constant at any one place and the changes between locations can be very precisely determined. The problem is to measure its absolute magnitude at some one selected place.

Our interest in the problem arises this way. In order to have a consistent set of units and standards in the various fields of science, each must be appropriately related to the arbitrary prototype standards of mass, length, time and temperature through an unbroken chain of measurement. The determination of g provides the transfer from these to force measurements and thence, for example, to the electrical standards and via them to our knowledge of the fundamental atomic constants, e, h, m, etc.

The unit of force follows from Newton's law

$$f = m\,a$$

as that force which will impart unit acceleration to unit mass. Now the attraction of the earth provides a convenient reproducible force acting upon every mass. Unfortunately, this force at the surface of the earth, where we are interested in it, is not unity on unit mass. If a mass is allowed to fall (accelerate), it does not accelerate with unit acceleration but with an acceleration g. However, if we measure carefully the acceleration g, we can then measure a force by means of a balance. We let the

force pull one arm of the balance and hang weights from the other arm until true balance is indicated. If m is the mass of the weights added, then the unknown force is given by

$$f = m g.$$

We thus see that g is a transfer constant enabling us to make force measurements in terms of our standards of mass, length and time, for the measurement of g is essentially a precise determination of how long (time) it takes a body to fall a given distance (length).

You may be thinking that it should be possible to arrange a force which would give unit acceleration to unit mass and use it for our standard. This could, of course, be done but no one has devised a system which will do it as precisely and reporducibly as the scheme which uses the attraction of the earth.

Now, our electrical standards are based upon the ampere and the ohm. To determine the ampere, in absolute units, we measure the force between two conductors carrying a current. Thus, g gets into the ampere. The ohm does not involve it, so we drop it from consideration here. Our measurements of many constants and in particular the atomic constants are done by means of electric and magnetic fields and hence involve the ampere, also unavoidably g.

It is indeed surprising to find that our presently accepted value of this important transfer constant g depends upon three "independent" measurements <u>all</u> using the method of the Kater reversible pendulum.

The results of these determinations are referred, by means of very precise transfer measurements, to one specific location Potsdam, Germany. They are shown in the table

Potsdam	1906	Kuhnen & Furtwangler	980.100
Dryden Revision	1942	Dryden (NBS)	980.088
Washington (NBS)	1936	Heyl & Cook	980.080
Teddington, England (NPL)	1939	Clark	980.084
	Mean of last three		980.084
	P.E. of mean	2 in 10^6	

This looks like very good agreement but attention should be called to the 1942 revision of the 1906 measurement. This shows that a later look at the same data brings a change of about 12 parts per million. Also, all the measurements are subject to the same possible systematic errors and so the measurements are not truly independent. In fact, study shows that a systematic error estimated to be as large as 15 ppm could be present.

Thus, the experiments show that the probable error for measurements of g by reversible pendulums is about 2 ppm. There is already some preliminary evidence based upon measurements by other methods that these measurements do in fact have an error of about 10 parts per million from the true value.

Here at NBS two of our scientists C. H. Page and D. R. Tate are now designing new experiments to get at the answer by methods which differ in principle from the older ones.

They will use a quite different type of pendulum and also will time a freely falling object, falling in vacuum. They are making every effort to design the experiment to eliminate known sources of error, to have each error subject to experimental estimation or below the desired limit of accuracy (about 1 part per million) and to take advantage of the use of statistical variation of parameters in the experiment itself. They are working closely in their work with our Statistical Engineering Section to get the benefit of their advice in the design phases of the experiment instead of waiting until the data is in as is all too often done.

Unless one has had an opportunity to participate in one of these precision measurements, it is difficult to understand the complexities that arise. In pendulums, the motions cause bending and stretching, minute temperature changes cause changes of length, wear changes the form of the bearings, even stray electric and magnetic fields cause significant perturbations. In the free fall experiment, mention of only one of many difficulties indicates the kind of factors that must be considered. One assumes that the laboratory is at rest on the earth during the time the object falls. This is not strictly true. Due to minor earthquakes, microseisms, the laboratory does not stay at rest with the precision needed. It is, therefore, necessary to set up a seismograph and record the microseisms. The free fall can then be made during quiet periods and corrections can be made for the motion of the laboratory during the fall. These motions may be as small as 40 millionths of an inch but they are still significant.

It is the consideration of the whole array of such errors and the design of experiment to take account of them that makes precision measurement such a fascinating science and one which depends very strongly upon proper design of experiment.

RECENT RESEARCH IN STATISTICAL PROBLEMS IN SUBJECTIVE TESTING[1,2]

Ralph Allan Bradley
Virginia Agricultural Experiment Station
of the Virginia Polytechnic Institute

1. INTRODUCTION. There is widespread interest in the design, conduct, and analysis of experiments involving the subjective opinions of samples and panels of individuals. Applications arise in food processing, photography, distilling and brewing, textile research, wood technology, petroleum products research, and in a host of other areas of research.

Problems, many of which at least have statistical aspects, arise in the selection of consumer samples and expert taste panels, in the training of panel members, in the design of experiments, in the development of scoring scales, and in the analysis and interpretation of experimental data. We shall present the results of recent research and illustrative examples on techniques that deal with the sensitivities of scoring scales, the variabilities of judges using scoring scales, the design of experiments with scoring scales, and the design of ranking experiments.

We shall not here discuss in any detail the selection or training of a taste panel, the selection of a consumer panel, or the development of a scoring scale. Some general discussion of the problems involved are given in the reference (Bradley [1953]) which has a large classified bibliography including papers on these subjects. Expert taste panels are usually selected through use of a system of triangle tests (a triangle test involves the selection of the odd sample from three samples of which two are identical). In the cited reference, we illustrate the use of sequential triangle tests. Hopkins and Gridgeman (1955) compare the sensitivities of paired and triad flavor intensity difference tests. Kramer (1955, 1956) has provided tables and discussions on the use of multiple matching systems for the selection of judges as an alternative to use of triangle tests. Procedures for the selection of a consumer panel should basically depend on sampling survey techniques and those used in opinion polls. In such studies it is well to keep the techniques simple and paired-sample preference tests are usually used along with a supplementary questionnaire. Ranking techniques in paired comparisons may be used in these surveys and the method is summarized in a subsequent section. There are many psychological aspects to the development of a scoring scale and we shall not discuss them here. When a scale is developed, the distributions of scores on the scale should be examined. Hopkins (1950) considered such distributions.

1. Presented at the 1956 Gordon Conference on Statistics in Chemistry and Chemical Engineering, New Hampton, N. H., August 23, 1956.

2. A report based largely on research sponsored by the Agricultural Research Service, U. S. D. A., under a Research and Marketing Act Contract, No. 12-14-100-126(20).

In the following sections we shall use the notation of the various basic reference papers rather than maintain a more uniform notation in this paper. This should permit the reader to more easily associate our examples with the theory in the references.

2. SENSITIVITY COMPARISONS. In the development of scoring scales and other experimental techniques, it is often desirable that two alternative methods be compared. Cochran (1943) discussed the comparison of different scales of measurement for experimental results and indicated where further research was required. We have provided means of comparing the sensitivities of similar experiments in two recent papers (Schumann and Bradley [1956], Bradley and Schumann [1956]). This recent research permits a test on the equality of the parameters of non-centrality of F-distributions associated with tests of treatment equality in two independent but parallel experiments containing the same set of treatments in identical experimental designs. The experiments may differ in the scoring scale used or in some other criterion of measurement that does not interact with treatments. Good experimental data to illustrate the method appeared as this paper was in preparation.

Kauman, Gottstein, and Lantican (1956) were interested in the quality evaluation of dried veneer. Two schemes were used to evaluate quality of sheets of veneer and they are designated as "numerical" and "subjective" although both were somewhat subjective. In the numerical scheme various types of degrade were listed with numerical scores for the severity of the degrade and weights were given for use in combining degrade scores to obtain a quality score. A quality rating of 50 in the numerical scheme was very bad and the maximum possible score; a quality rating of 0 was excellent and indicates a sheet free from degrade. In the subjective scheme "quality ratings" were assigned on a 0-8 scale with 0, excellent and 8, very bad. Twenty selected sheets of veneer were evaluated by three observers, twice with each scheme, and repeat observations were spaced by several days with the order of presentation of the sheets changed. The complete tables of scores are given in the reference; we repeat the analyses of variance in Table 1.

Table 1

Analyses of Variance for Quality Ratings*

Factor	Degrees of freedom	Numerical scheme		Subjective scheme	
		Sum of squares	Mean square	Sum of squares	Mean square
Sheets (S)	19	12826.16	675.1	336.90	17.73
Observers (O)	2	170.72	85.36	3.70	1.852
Repetitions (R)	3	168.13	56.04	0.61	0.2042
Interaction (SO)	38	823.61	21.67	30.12	0.7928
Error (SR)	57	595.37	10.45	22.13	0.3884

* A reproduction of part of Table 6, Kauman, Gottstein, and Lantican (1956), page 148.

While the authors of the cited reference properly considered Model II of the analysis of variance and estimated variance components, we shall illustrate how to apply a test of the sensitivities of the two experiments conditional on the observers and samples actually used in the experiments and assume Model I of analysis of variance with "fixed" effects. Under these conditions, the expected value of the mean square for sheets is

$$(1.2) \qquad E[M.S.(S)] = \sigma^2 + k\sum_{i=1}^{t}\tau_i^2/(t-1)$$

is general for t sheets and k observations on each sheet. τ_i is the "effect" of sheet i, i = 1, ...,t. In the examples,

$$(2.2) \qquad E[M.S.(S)] = \sigma^2 + 6\sum_{i=1}^{20}\tau_i^2/19.$$

σ^2 is the expectation of the error mean square in both (1.2) and (2.2). The parameter of non-centrality of the F-test for sheets is, in general,

$$(3.2) \qquad \lambda = k\sum_{i=1}^{t}\tau_i^2/2\sigma^2$$

and, in the examples,

$$(4.2) \qquad \lambda = 6\sum_{i=1}^{20}\tau_i^2/2\sigma^2$$

when the F-density is written

$$(5.2) \qquad f(F) = (a/b)^a[B(a,b)]^{-1}e^{-\lambda}F^{a-1}(1+aF/b)^{-(a+b)} \cdot {}_1F_1\left[a+b,\ a,\ a\lambda F/b\,(1+aF/b)\right],\ 0 \leq F \leq \infty$$

where F has 2a and 2b degrees of freedom, ${}_1F_1$ is the confluent hypergeometric series, and B represents the beta function. It is seen at once that λ is a parameter expressing the magnitudes of treatment effects in a scale in terms of the experimental error associated with the scale. λ is the appropriate parameter to measure the sensitivity of a scale. We shall test the hypothesis, $H_0: \lambda_1 = \lambda_2$, against the alternative, $H_a: \lambda_1 \neq \lambda_2$, using the subscripts 1, for the numerical scheme, and 2, for the subjective scheme.

To apply the test, we compute the two F-ratios with 19 and 57 degrees of freedom (Now a = 9.5, b = 28.5.) and obtain

$$F_1 = 64.60 \qquad \text{and} \qquad F_2 = 45.65.$$

The statistic used is

$$(6.2) \qquad w = F_1/F_2 = 64.60/45.65 = 1.42.$$

The distribution of w under H_0 depends on $\lambda = \lambda_1 = \lambda_2$ which in general is unknown. In practice it is clear that the test is not very sensitive to small changes in λ and we in fact estimate λ from the data using

(7.2) $$\hat{\lambda}_i = a(F_i - 1), \quad i = 1, 2.$$

In the examples,

$$\hat{\lambda}_1 = 9.5(64.60-1) = 604.2$$

and

$$\hat{\lambda}_2 = 9.5(45.65-1) = 424.2.$$

We take $\hat{\lambda}$ to be the average of $\hat{\lambda}_1$ and $\hat{\lambda}_2$,

(8.2) $$\hat{\lambda} = \tfrac{1}{2}(604.2 + 424.2) = 514.2.$$

A table of values w_0 such that $P(w > w_0 | H_0) = 0.05$ is given by Bradley and Schumann in the cited references. To enter this table, one requires

(9.2) $$a' = (a+\lambda)^2/(a+2\lambda) = (9.5+514.2)^2/(9.5+1028.4)$$
$$= 264.2$$

and $b = 28.5$. The table is symmetric in the sense that $w_0(a',b) = w_0(b,a')$ and we obtain $w_0 \approx 1.85$ by consulting the table. Now H_a, as postulated, is two-sided and hence the significance level being used is 0.10. w in (6.2) does not exceed w_0 and consequently we do not reject H_0 at the 10% level of significance. We are in accord with the authors (Kauman et al.) who state "the present experiment has shown that the subjective evaluation can yield results of an accuracy approaching that of the numerical scheme, although the accuracy of the latter was slightly superior".

The theory of the test of sensitivity is given in detail by Schumann and Bradley (1956) and other applications are given by Bradley and Schumann (1956). As a somewhat different application, the method may also be used to compare values of R^2, the square of the multiple correlation coefficient, for two similar but independent regression studies based on the usual regression model. The theory involves an approximation which appears to be good. The distribution of w should come from the joint distribution of two non-central variance-ratios with equal pairs of degrees of freedom and equal parameters of non-centrality. What was done was to approximate to the non-central F-distributions using central F-distributions and to obtain the distribution of w taking w to be the ratio of two independent central F-variates.

Applications are limited since a table is only available for a one-sided 5% level test. Schumann is preparing additional tables.

3. JUDGE VARIABILITY AND JUDGE COMPARISONS. When items are scored in subjective experimentation, there is no knowledge of the "true worth" of the sample in the units of the scoring scale. It is then difficult to assess the judging ability of a judge. Russell and Bradley (1956) have provided means of estimating the variability of a judge in terms of the deviations of his scores for an item from those of the remaining judges but permitting a judge a possible constant bias in his assignment of scores. Similar procedures were considered by Grubbs (1948) and Ehrenberg (1950) and they obtained the same estimators from somewhat different demonstrations but did not develop the test procedures illustrated below.

Consider a two way classification with t items or treatments and r judges. The model with fixed effects is

$$(1.3)\quad y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij},\ i = 1,\ldots,t,\ j = 1,\ldots,r$$

where y_{ij} is the score assigned by the j^{th} judge to the i^{th} item, μ is the grand mean, the average level of judging, τ_i is the effect of the i^{th} item, β_j is the effect (or bias) of the j^{th} judge, and ε_{ij} are independent normal variates with zero means. Contrary to the usual model of analysis of variance, we admit the possibility of heterogeneous error variances in the sense that

$$(2.3)\quad E(\varepsilon_{ij}^2) = \sigma_j^2,$$

σ_j^2 is the variance of the j^{th} judge and is to be estimated.

The estimator of σ_j^2 to be used is

$$(3.3)\quad \hat{\sigma}_j^2 = \frac{rG_j}{(t-1)(r-2)} - \frac{E}{(t-1)(r-1)(r-2)}$$

where

$$(4.3)\quad G_j = \sum_{i=1}^{t} (y_{ij}-y_{i.}-y_{.j}+y_{..})^2$$

and

$$(5.3)\quad E = \sum_{i=1}^{t}\sum_{j=1}^{r} (y_{ij}-y_{i.}-y_{.j}+y_{..})^2,$$

the latter being the error sum of squares from the analysis of variance of the two-way classification. $\hat{\sigma}_j^2$ is an unbiased estimator of σ_j^2 but, like an estimate of a variance component, may occasionally be negative. In (4.3) and (5.3), $y_{i.}$ is the average of scores for treatment i, $y_{.j}$ is the average of scores assigned by the j^{th} judge, and $y_{..}$ is the average of all scores. The requirement that ε_{ij} in (1.3) be normal is only met approximately in use of a discrete scoring scale but does not affect the estimation of σ_j^2. In later paragraphs of this section, we shall assume that departures from non-normality do not seriously affect our test procedures.

We shall again illustrate this work using the data of Kauman et al. The detailed example is for Test 1 using the subjective scheme. Scores are listed in Table 2. In Table 3 we show values of $(y_{i.} + y_{.j} - y_{..})$ obtained by first writing down the marginal entries and then computing the required table entries. In Table 4 we have the residuals, $(y_{ij} - y_{i.} - y_{.j} + y_{..})$, obtained by subtracting entries in Table 3 from corresponding entries in Table 2. Values of G_j and E are given in the lower margin of Table 4 and are obtained by accumulating the squares of entries in the columns above as required in view of (4.3). $\sum_{j=1}^{r} G_j$ was so obtained. The values of $\hat{\sigma}_j^2$

$$E = \sum_{j=1}^{r} G_j$$

computed using (3.3) are listed in Table 5 along with those for the other three tests of Kauman et al. To illustrate the computations, we use observer A and obtain

$$\hat{\sigma}_1^2 = \frac{3(8.36)}{(19)(1)} - \frac{22.50}{(19)(2)(1)} = 0.72.$$

Certain checks on the computation are possible. The residuals in Table 4 have row and column totals that are zero except for rounding. Also, as already noted, $\sum_{j=1}^{r} G_j = E$ and E will usually have been obtained directly from the analysis of variance. A final check follows from the fact that

$$E = \frac{(t-1)(r-1)}{r} \sum_{j=1}^{r} \hat{\sigma}_j^2. \tag{6.3}$$

In the example,

$$\frac{(t-1)(r-1)}{r} \sum_{j=1}^{r} \hat{\sigma}_j^2 = \frac{(19)(2)}{3} [(0.72) + (0.53) + (0.53)] = 22.55.$$

A test of homogeneity of variances is possible only when r = 3. The only situation wherein the estimators $\hat{\sigma}_j^2$ of σ_i^2 are maximum likelihood estimators is when r = 3 and then an approximate test may be made. Consider the hypothesis,

$$H_o: \sigma_1^2 = \sigma_2^2 = \sigma_3^2,$$

and the alternative,

$$H_a: \sigma_j^2 \neq \sigma_k^2 \text{ for some } j \text{ and } k,\ j, k = 1, 2, 3.$$

The likelihood ratio test statistic, distributed approximately as χ^2-variate with 2 degrees of freedom for large samples, is

$$\chi_2^2 = -(2.3026)(t-1)[2 \log(t-1) + \log(\hat{\sigma}_1^2\hat{\sigma}_2^2 + \hat{\sigma}_1^2\hat{\sigma}_3^2 + \hat{\sigma}_2^2\hat{\sigma}_3^2) - 2\log E + \log 4/3] \tag{7.3}$$

$$= -(2.3026)(19)[2 \log 19 + \log \{(0.72)(0.53) + (0.72)(0.53) + (0.53)(0.53)\} - 2 \log 22.50 + \log 4/3] = 0.14.$$

Table 2

Quality Ratings for the Subjective Quality Evaluation Test 1*

Sheet No.	Observers A	B	C	$y_{i.}$
1	3	3	3	3.00
2	7	5	7	6.33
3	6	5	5	5.33
4	7	8	7	7.33
5	1	2	3	2.00
6	5	5	5	5.00
7	5	6	5	5.33
8	3	5	4	4.00
9	4.5	5	5	4.83
10	6	7	7	6.67
11	5	4	4	4.33
12	8	7	8	7.67
13	5	7	5	5.67
14	1	2	2	1.67
15	7	7	7	7.00
16	1	3	4	2.67
17	4	3	3	3.33
18	6	6	5	5.67
19	5	5	7	5.67
20	3	3	2	2.67
$y_{.j}$	4.62	4.90	4.90	$y_{..}$ 4.81

Table 3

Values of $(y_{i.}+y_{.j}-y_{..})$ for the Subjective Quality Evaluation Test 1

Sheet No.	Observers A	B	C	$y_{i.}$
1	2.81	3.09	3.09	3.00
2	6.14	6.42	6.42	6.33
3	5.14	5.42	5.42	5.33
4	7.14	7.42	7.42	7.33
5	1.81	2.09	2.09	2.00
6	4.81	5.09	5.09	5.00
7	5.14	5.42	5.42	5.33
8	3.81	4.09	4.09	4.00
9	4.64	4.92	4.92	4.83
10	6.48	6.76	6.76	6.67
11	4.14	4.42	4.42	4.33
12	7.48	7.76	7.76	7.67
13	5.48	5.76	5.76	5.67
14	1.48	1.76	1.76	1.67
15	6.81	7.09	7.09	7.00
16	2.48	2.76	2.76	2.67
17	3.14	3.42	3.42	3.33
18	5.48	5.76	5.76	5.67
19	5.48	5.76	5.76	5.67
20	2.48	2.76	2.76	2.67
$y_{.j}$	4.62	4.90	4.90	$y_{..}$ 4.81

* From Table 3, Kauman, Gottstein, and Lantican (1956), page 135.

Table 4

Values of $(y_{ij}-y_{i.}-y_{.j}+y_{..})$ for the Subjective Quality Evaluation, Test 1

Sheet No.	Observers		
	A	B	C
1	0.19	-0.09	-0.09
2	0.86	-1.42	0.58
3	0.86	-0.42	-0.42
4	-0.14	0.58	-0.42
5	-0.81	-0.09	-0.91
6	0.19	-0.09	-0.09
7	-0.14	0.58	-0.42
8	-0.81	0.91	-0.09
9	-0.14	0.08	0.08
10	-0.48	0.24	0.24
11	0.86	-0.42	-0.42
12	0.52	-0.76	0.24
13	-0.48	1.24	-0.76
14	-0.48	0.24	0.24
15	0.19	-0.09	-0.09
16	-1.48	0.24	1.24
17	0.86	-0.42	-0.42
18	0.52	0.24	-0.76
19	-0.48	-0.76	1.24
20	0.52	0.24	-0.76
G_j	8.36	7.07	7.07

$E = 22.50$

The multiplier, 2.3026, in (7.3) is included so that common logarithms may be used in the computation of χ^2. The small value of χ^2 indicates that the observers may be taken to have homogeneous variances.

In Table 5, we have included values of χ^2_2 for all four tests and show also values of $\hat{\sigma}^2$, the error mean square from the analysis of variance. Note that only in one of the numerical tests was χ^2_2 significant at the 5% level of significance. The estimates of variance in the numerical scheme are considerably larger than in the subjective scheme. This does not of course suggest a preference for the subjective scheme but is merely a result of the scales used in the scoring methods. The appropriate method of comparing the scales is the one given in the preceding section.

Table 5

Estimates of Variance and χ^2_2 to Test for Homogeneity of Observer Variances for All Four Tests of Kauman et al.

Tests	Observer Variances, $\hat{\sigma}^2_j$			Error Mean Square, $\hat{\sigma}^2$	χ^2_2
	A	B	C		
Subjective Test 1	0.72	0.53	0.53	0.59	0.14
Subjective Test 2	0.46	0.58	0.74	0.59	0.26
Numerical Test 1	4.58	10.79	33.02	16.13	6.42*
Numerical Test 2	2.19	26.40	21.69	16.84	4.14

Observer A was the only observer with previous experience in judging veneer except for brief training sessions before the experiment began. Another test, and this is an exact test, is possible. Consider the null hypothesis

$$H_o: \sigma_1^2 = \sigma^2, \text{ given } \sigma_2^2 = \ldots = \sigma_r^2 = \sigma^2,$$

and the alternative,

$$H_a: \sigma_1^2 < \sigma^2, \text{ given } \sigma_2^2 = \ldots = \sigma_r^2 = \sigma^2.$$

the statistic used is

$$(8.3) \qquad F = \frac{r(r-2)G_1}{(r-1)E-rG_1}$$

with (t-1) and (t-1)(r-2) degrees of freedom. H_a may have either of the possible one-sided forms or be two-sided. For the form of H_a shown, small values of F are significant. In the example, there is no point in testing H_o versus H_a in this test in view of the homogeneity of variances demonstrated above. However, we shall proceed in order to illustrate the method.

"adaption", the effect of the presence of one treatment on another in the same incomplete block. A doubly balanced incomplete block design is one, which in addition to being balanced, has all triplets of treatments appearing in incomplete blocks an equal number of times. The use of doubly balanced incomplete block designs permitted easy evaluation of the additional parameters inserted in the linear model. Calvin's model is

$$(2.4) \qquad y_{hi} = n_{hi}\left(\mu + \beta_h + \tau_i + \sum_{\substack{j \\ j \neq i}} n_{hj} m_{ij} \alpha_{ij} + \varepsilon_{hi}\right)$$

where

y_{hi} is an observation on treatment i in block h,

n_{hi} = 1 if treatment i occurs in block h

= 0 otherwise,

μ represents the average level of scoring,

β_h represents the effect of block h (perhaps due to the taster doing the scoring, the time of day, etc.),

τ_i is the effect of treatment i,

m_{ij} = 1 if i < j

= -1 if j < i,

α_{ij} is the effect of the presence of treatment j on treatment i ($\alpha_{ij} = -\alpha_{ji}$), and

ε_{hi} is equivalent to e_{ijk} in (1.4) above.

Calvin called the effects measured by α_{ij}, the correlation effects. We shall not give examples of analyses using either the Scheffé or the Calvin designs here but instead refer the reader to the references for such examples.

Factorial treatment combinations are often required in subjective testing, for food samples may result from a variety of process changes in their manufacture as may photographic samples, dye samples and the like. Means of incorporating factorial treatments in incomplete block designs are then required. That this may be done in balanced incomplete block designs seems to be well known although we have not found a direct reference. Kramer and Bradley (1956, 1956a) have shown how to use factorials in group-divisible, two-associate class, partially balanced, incomplete block designs. We shall give an example here and note that additional examples are given in the references along with the theory. We use only an intra-block analysis of variance; Walpole at the Virginia Polytechnic Institute is considering inter-block analyses. Kramer is also considering extensions to other types of two-associate class, partially balanced, incomplete block designs.

A group-divisible, two-associate class, partially balanced, incomplete block design has design parameters as follows:

v: the number of treatments or varieties,
r: the number of observations on each treatment,
k: the number of units in an incomplete block,
b: the number of incomplete blocks,
m: the number of groups,
n: the number of treatments in a group, $v = mn$,

where treatments in the same group are first associates and treatments are not in the same group are second associates.

λ_1: the number of times two first associate treatments appear together in incomplete blocks, and

λ_2: the number of times two second associate treatments appear together in incomplete blocks. For these designs, the treatments may be given in an m by n rectangular association scheme. Bose, Clatworthy, and Shrikhande (1954) have catalogued all known designs of this class with block size, $3 \leq k \leq 10$ and $r \leq 10$. We consider an example with $v = 8$, $r = 3$, $k = 3$, $b = 8$, $m = 4$, $n = 2$, $\lambda_1 = 0$, $\lambda_2 = 1$ designated as Design R5 of the catalogue. This is a made-up example as no data were available.

For the example, we have the basic association scheme of Table 6 where treatments in the same row are first associates and we use a double subscript notation to designate treatments and the symbol $V_{\cdot}$. In Table 7 we show the association scheme for a 4 by 2 factorial with an A-factor at four levels and a C-factor at two levels. The design lay-out, observations, block totals $B_{\cdot}$ and grand total G, are given in Table 8. The treatments in Table 8 are associated with the factorials through the correspondence of items in Tables 6 and 7.

Table 6

Association Scheme for 8 Treatments

V_{11}	V_{12}
V_{21}	V_{22}
V_{31}	V_{32}
V_{41}	V_{42}

Table 7

Association Scheme for the 4x2 Factorial

A_1C_1	A_1C_2
A_2C_1	A_2C_2
A_3C_1	A_3C_2
A_4C_1	A_4C_2

Table 8

Design and Observations for the Eight Treatments

Blocks	Observations			B_s
1	V_{11} 35	V_{21} 24	V_{41} 39	98
2	V_{21} 42	V_{31} 45	V_{12} 48	135
3	V_{31} 13	V_{41} 15	V_{22} 17	45
4	V_{41} 19	V_{12} 22	V_{32} 25	66
5	V_{12} 28	V_{22} 30	V_{42} 31	89
6	V_{22} 51	V_{32} 52	V_{11} 54	157
7	V_{32} 60	V_{42} 65	V_{21} 67	192
8	V_{42} 54	V_{11} 57	V_{31} 58	169
			Total G =	951

The basic analysis of variance without consideration of the factorial effects is straight-forward. The total sum of squares and the block sum of squares are computed in the usual way. We find it useful to compute the adjusted treatment sum of squares from the estimates of treatment effects. The linear model is

$$(3.4) \qquad y_{ijs} = \mu + \tau_{ij} + \beta_s + \varepsilon_{ijs}$$

where y_{ijs} is the observation on V_{ij} if that treatment is in block s, μ is the over-all average, τ_{ij} is the effect of V_{ij}, β_s is the effect of block s, and ε_{ijs} is the error variate as described after (1.4). If t_{ij} is the estimator of τ_{ij}, in general,

$$(4.4) \quad t_{ij} = \left[k v\lambda_2 T_{ij} - k(\lambda_2-\lambda_1)\Sigma_j T_{ij} - v\lambda_2 B_{ij\cdot} + (\lambda_2-\lambda_1)\Sigma_j B_{ij\cdot}\right] / v\lambda_2(\lambda_1+rk-r)$$

and, in the example,

$$(5.4) \quad t_{ij} = \frac{1}{2}T_{ij} - \frac{1}{16}\Sigma_j T_{ij} - \frac{1}{6}B_{ij\cdot} + \frac{1}{48}\Sigma_j B_{ij\cdot}.$$

T_{ij} is the total for V_{ij}; $B_{ij\cdot}$ is the total of block totals of those blocks containing V_{ij}. Values of T_{ij}, $\Sigma_j T_{ij}$, $B_{ij\cdot}$, $\Sigma_j B_{ij\cdot}$, and t_{ij} are given in Table 9 in positions corresponding to the array of Table 6. In addition, in Table 9, we show the totals, $t_{i\cdot} = \Sigma_j t_{ij}$, $t_{\cdot j} = \Sigma_i t_{ij}$, and the averages, $\bar{t}_{i\cdot} = t_{i\cdot}/n$ and $\bar{t}_{\cdot j} = t_{\cdot j}/m$. The adjusted treatment sum of squares may in general be written as

$$(6.4) \quad \text{Adj.Treat.S.S.} = \frac{(\lambda_1+rk-r)}{k}\Sigma\Sigma_{ij} t_{ij}^2 + \frac{(\lambda_2-\lambda_1)}{k}\Sigma_i t_{i\cdot}^2$$

and here becomes

$$(7.4) \quad \text{Adj.Treat.S.S.} = 2\Sigma\Sigma_{ij} t_{ij}^2 + \frac{1}{3}\Sigma_i t_{i\cdot}^2 = 49.54.$$

DK4

OFFICE OF ORDNANCE RESEARCH, U. S. ARMY
BOX CM, DUKE STATION
DURHAM, NORTH CAROLINA

IN REPLY
REFER TO:
ORDOR-MD

22 April 1958

Dr. I. R. Hershner, Jr.
Research and Development Field
Office of the Chief of Research and Development
Fort Belvoir, Virginia

Dear Ray:

Following the Second Conference on the Design of Experiments in Army Research, Development and Testing, copies of most of the papers presented at the meeting were collected from the authors. This group of articles has now been published in the Proceedings of the afore mentioned conference, and we are inclosing a copy for your use.

Sincerely yours,

F. G. DRESSEL
Assistant
Mathematical Sciences Division

1 Incl
Proceedings

Table 9

Values of T_{ij}, $\Sigma_j T_{ij}$, $B_{ij.}$, $\Sigma_j B_{ij.}$, and t_{ij}

T_{ij}		$\Sigma_j T_{ij}$	$B_{ij.}$		$\Sigma_j B_{ij.}$	t_{ij}		Totals $t_{i.}$	Averages $\bar{t}_{i.}$
146	98	244	424	290	714	1.958	0.292	2.250	1.125
133	98	231	425	291	716	-3.854	0.979	-2.875	-1.438
116	137	253	349	415	764	-0.063	-0.562	-0.625	-0.313
73	150	223	209	450	659	1.458	-0.208	1.250	0.625
			Totals $t_{.j}$			-0.501	0.501	0.000	0.000
			Averages $\bar{t}_{.j}$			-0.125	0.125		

To complete the basic analysis, we have

$$(8.4)\quad \text{Unadj.Block S.S.} = \sum_{s=1}^{b} B_s^2/k - G^2/rv = 6384.95,$$

$$(9.4)\quad \text{Total S.S.} = \sum_{ijs}\sum\sum y_{ijs}^2 - G^2/rv = 6593.62,$$

and the error sum of squares is obtained by subtraction,

$$(10.4)\quad \text{Error S.S.} = \text{Total S.S.} - \text{Unadj.Block S.S.} - \text{Adj.Treat. S.S.}$$
$$= 159.13.$$

Degrees of freedom are: Treatments, $(v-1) = 7$; Blocks, $(b-1) = 7$; Error, $[v(r-1)-b+1] = 9$; Total, $(rv-1) = 23$. The analysis of variance is given in Table 13.

To consider the analysis for the 4 by 2 factorial of Table 7, we need only partition the adjusted treatment sum of squares into adjusted sums of squares for A-factor, C-factor, and AC-interaction. This is easily done and the basic formulas are

$$(11.4)\quad \text{Adj. A-factor S.S.} = (nK_1+n^2K_2)\Sigma_i \bar{t}_{i.}^2,$$

$$(12.4)\quad \text{Adj. C-factor S.S.} = mK_1\Sigma_j \bar{t}_{.j}^2,$$

(13.4) Adj. AC-Interaction S.S. $= K_1 \sum\sum_{ij} (t_{ij} - \bar{t}_{i.} - \bar{t}_{.j})^2$

where $K_1 = (\lambda_1 + rk - r)/k$, $K_2 = (\lambda_2 - \lambda_1)/k$. Usually we compute

Adj. AC-Interaction S.S. = Adj.Treat.S.S. - Adj. A-factor S.S.
- Adj. C-factor S.S.

In the example, $K_1 = 2$, $K_2 = 1/3$, $m = 4$, and $n = 2$. Then,

(14.4) Adj. A-factor S.S. $= \frac{16}{3} \sum_i \bar{t}_{i.}^2 = 20.37$,

(15.4) Adj. C-factor S.S. $= 8 \sum_j \bar{t}_{.j}^2 = 0.25$,

and

(16.4) Adj. AC-Interaction S.S. $= 2 \sum\sum_{ij} (t_{ij} - \bar{t}_{i.} - \bar{t}_{.j})^2 = 28.92$.

Single degree of freedom comparisons may be used. Consider linear, quadratic, and cubic trends over the levels of the A-factor and their interactions with the C-factor. This is done in much the usual way except that additional and different multipliers are required for components of the A-factor, C-factor and AC-interaction sums of squares. The method will be evident from Table 10 but to illustrate we consider the Linear A-component. The linear contrast for Linear A is

$$L(\text{lin.A}) = -3(1.958) - 3(0.292) + \ldots + 3(-0.208) = -0.750.$$

The sum of squared coefficients is

$$\Delta(\text{lin.A}) = (-3)^2 + (-3)^2 + \ldots + (3)^2 = 40.$$

The multiplier is, in general for a component of the A-factor,

$$M(\text{lin.A}) = (nK_1 + n^2K_2)/n = 8/3;$$

the adjusted sum of squares for the linear A-factor component is

$$\text{Adj.S.S. (Lin.A)} = \frac{[L(\text{lin.A})]^2}{\Delta(\text{lin.A})} \cdot M(\text{lin.A}) = \frac{(-0.750)^2(8/3)}{40} = 0.04.$$

In general, the multiplier for a component of the C-factor and for a component of the AC-interaction is K_1 itself.

Now we are not restricted to a two-factor factorial but in general may have several factors with levels $m_1, \ldots, m_p$ and $n_1, \ldots, n_q$ so long as $\sum_{i=1}^{p} m_i = m$ and $\sum_{j=1}^{q} n_j = n$. Suppose the A-factor were in fact a 2x2 factorial itself. If we designate these new factors as N and P and associate them with the association schemes of Tables 6 and 7 as given in Table 11, we can analyze the experiment as a 2^2x2 factorial subdividing the adjusted treatment sum of squares as in Table 12.

Table 10

Trend Components for the 4 by 2 Factorial

	Treatment Effects Estimates											
	t_{11}	t_{12}	t_{21}	t_{22}	t_{31}	t_{32}	t_{41}	t_{42}	Linear Contrast	Sum of Squared Coefficients	Multiplier	Adj. S.S.
	1.958	0.292	-3.854	0.979	-0.063	-0.562	1.458	-0.208				
Comparison	Comparison Coefficients											
Linear A	-3	-3	-1	-1	+1	+1	+3	+3	-0.750	40	8/3	0.04
Quad. A	+1	+1	-1	-1	-1	-1	+1	+1	7.000	8	8/3	16.33
Cubic A	-1	-1	+3	+3	-3	-3	+1	+1	-7.750	40	8/3	4.00
C	-1	+1	-1	+1	-1	+1	-1	+1	1.000	8	2	0.25
Lin.AxC	+3	-3	+1	-1	-1	+1	-3	+3	-5.333	40	2	1.42
Quad.AxC	-1	+1	+1	-1	+1	-1	-1	+1	-7.667	8	2	14.69
Cub.AxC	+1	-1	-3	+3	+3	-3	-1	+1	16.000	40	2	12.80
											Total	49.53

Table 11

Association Scheme for the $2^2 \times 2$ Factorial

$N_1P_1C_1$	$N_1P_1C_2$
$N_1P_2C_1$	$N_1P_2C_2$
$N_2P_1C_1$	$N_2P_1C_2$
$N_2P_2C_1$	$N_2P_2C_2$

The analysis of variance for the various breakdowns of the experimental data considered is given in Table 13.

We believe that these designs with factorial treatment combinations offer a useful aid in subjective experimentation. The analyses are reasonably simple and straight-forward and out of the many such designs catalogued it should be easy to select one appropriate for the planned research. Other applications in many fields of experimentation should be forthcoming.

Table 12

Comparisons for the 2^2 by 2 Factorial

	Treatment Effects Estimates											
	t_{11}	t_{12}	t_{21}	t_{22}	t_{31}	t_{32}	t_{41}	t_{42}	Linear Contrast	Sum of Squared Coefficients	Multiplier	Adj. S.S.
	1.958	0.292	-3.854	0.979	-0.063	-0.562	1.458	-0.208				
Comparison	Comparison Coefficients											
N	-1	-1	-1	-1	+1	+1	+1	+1	1.250	8	8/3	0.52
P	-1	-1	+1	+1	-1	-1	+1	+1	-3.250	8	8/3	3.52
NP	+1	+1	-1	-1	-1	-1	+1	+1	7.000	8	8/3	16.33
C	-1	+1	-1	+1	-1	+1	-1	+1	1.000	8	2	0.25
NC	+1	-1	+1	-1	-1	+1	-1	+1	-5.333	8	2	7.11
PC	+1	-1	-1	+1	+1	-1	-1	+1	5.333	8	2	7.11
NPC	-1	+1	+1	-1	+1	-1	-1	+1	-7.667	8	2	14.69
											Total	49.53

Table 13

Intra-Block Analysis of Variance for the Illustrative Experiment

Source	d.f.	S.S.	M.S.
Treat.(adj.)	7	49.54	7.08
Subdivision for 4 by 2 Factorial			
A-factor(adj.)	3	20.37	6.79
C-factor(adj.)	1	0.25	0.25
AC-interaction(adj.)	3	28.92	9.64
Subtotals	7	49.54	
Subdivision for Trends in 4 by 2 Factorial			
Linear A(adj.)	1	0.04	0.04
Quad. A(adj.)	1	16.33	16.33
Cubic A(adj.)	1	4.00	4.00
C-factor(adj.)	1	0.25	0.25
Linear A by C(adj.)	1	1.42	1.42
Quad. A by C(adj.)	1	14.69	14.69
Cubic A by C(adj.)	1	12.80	12.80
Subtotals	7	49.53	
Subdivision for 2^2 by 2 Factorial			
N (adj.)	1	0.52	0.52
P (adj.)	1	3.52	3.52
NP (adj.)	1	16.33	16.33
C (adj.)	1	0.25	0.25
NC (adj.)	1	7.11	7.11
PC (adj.)	1	7.11	7.11
NPC (adj.)	1	14.69	14.69
Subtotals	7	49.53	
Blocks (unadj.)	7	6384.95	912.14
Error	9	159.13	17.68
Total	23	6593.62	

5. RANKING METHODS FOR SUBJECTIVE TESTING. We have discussed statistical methods for subjective testing for use with scoring scales up to this point. It is our opinion, and one that is not easy to prove or disprove, that in many experimental situations it is easier and more efficient to use ranking methods rather than scoring methods. Any loss in efficiency due to ranking, if indeed there is such loss of efficiency, may be offset by increased ease and speed of experimentation which permits use of increased sample sizes for the same time of experimentation. As we see it, the disadvantages of using ranking methods is that such methods are not fully developed. Experimental designs that permit use of factorials in incomplete blocks are not directly available unless one is willing to use analysis of variance on ranks transformed to scores through use of Table XX of Fisher and Yates (1948). Similarly, except for the method of paired comparisons (which is widely applicable), we do not have well developed ranking methods for use in incomplete blocks unless transformation is again used. We shall briefly review, but not discuss in detail, the method of paired comparisons introduced by Bradley and Terry (1952) and Terry, Bradley and Davis (1952) and the method of concordance for ranking in balanced incomplete block designs presented by Durbin (1951).

Consider t treatments in n repetitions of the possible $t(t-1)/2$ paired treatment comparisons. The basic model for the method of paired comparisons assumes the existence of parameters, $\pi_1,\ldots,\pi_t$, $\pi_i \geq 0$, $\Sigma_i \pi_i = 1$ such that, if X_i is an observation on treatment i and X_j, on treatment j, the probability that $X_i < X_j$, treatment i receives rank 1 and treatment j receives rank 2, treatment i is preferred to treatment j, is

$$P(X_i < X_j) = \pi_i/(\pi_i+\pi_j). \tag{1.5}$$

Methods of maximum likelihood are used to obtain estimators p_i of π_i. These estimators are obtained by solution of (t+1) simultaneous (but not independent) equations

$$\frac{a_i}{p_i} - \sum_{j \neq i} \frac{n}{p_i+p_j} = 0, \quad i = 1,\ldots,t, \tag{2.5}$$

$$\sum_i p_i = 1 \tag{3.5}$$

where $a_i = 2n(t-1)-\Sigma r_i$, Σr_i is the total sum of ranks for treatment i, and a_i is essentially the number of times treatment i was given first choice. Difficulties in application stem from the problem of solving equations (2.5) and (3.5). Iterative methods have been suggested and tables of values of Σr_i and p_i are given in the first reference cited on this subject and by Bradley (1954). Recently Dykstra (1956) has provided easy means of obtaining good approximations to the solutions of these equations and, if his approximations are used as first estimates of the solution, at most one or two iterations are required to obtain the solution with desired accuracy. When the estimators are obtained, a test of the hypothesis of treatment equality,

$$H_o: \quad \pi_i = 1/t, \quad i = 1,\ldots,t,$$

against the general alternative,

$$H_a: \pi_i \neq 1/t \text{ for some } i,$$

is based on the statistic,

$$(4.5) \qquad B_1 = \sum_{i<j} \log(p_i+p_j) - \sum_i a_i \log p_i$$

and

$$(5.5) \qquad -2 \ln \lambda = (2.3026)\left[nt(t-1)\log 2 - 2B_1\right],$$

the latter of which has approximately the χ^2-distribution with $(t-1)$ degrees of freedom.

The method of paired comparisons has been further developed. The experiment may be performed in groups of repetitions (by judges, days, etc.) and a test of group by treatment interaction is possible. A test for the appropriateness of the model is discussed by Bradley (1954a) and tests on the model using extensive experimental data were made by Hopkins (1954). The properties of the method and power comparisons are the subject of a paper by Bradley (1955) and Abelson and Bradley (1954) attempted to incorporate factorial arrangements of treatments into paired comparisons. Algebraic difficulties essentially prohibit the use of factorials in practise. Wilkinson (1956) in a thesis considered the use of our model for paired comparisons in certain designs of Bose with blocks of size two.

Durbin generalized the method of concordance, previously available for paired comparisons and randomized block designs [Kendall (1948)]. Durbin supposed that n objects are presented in blocks of size k with each object ranked m times in the experiment. λ is the number of blocks in which any particular pair of treatments or objects occur, $\lambda = m(k-1)/(n-1)$. The coefficient of concordance, in the absence of tied ranks is

$$(6.5) \qquad W = \frac{12 \sum_{j=1}^{n} x_j^2 - 3m^2 n(k+1)^2}{\lambda^2 n(n^2-1)}$$

where x_j is the total sum of ranks for the j^{th} object. A test for independence among the m rankings of an object, essentially a test for treatment effects, is made by computing

$$(7.5) \qquad F = \frac{\left[\frac{\lambda(n+1)}{(k+1)} - 1\right] W}{1-W}$$

and taking this statistic to have approximately the F-distribution of analysis of variance with ν_1 and ν_2 degrees of freedom where

$$(8.5) \qquad \nu_1 = \frac{mn\left[1 - \frac{(k+1)}{\lambda(n+1)}\right]}{\left[\frac{nm}{(n-1)} - \frac{k}{(k-1)}\right]} - \frac{2(k+1)}{\lambda(n+1)}$$

and

$$(9.5) \qquad \nu_2 = \left[\frac{\lambda(n+1)}{(k+1)} - 1\right]\nu_1.$$

ν_1 and ν_2 may not be integers but interpolation in F-tables is possible. A numerical example is given by Durbin and a somewhat large example is given by Bradley (1955a).

6. DISCUSSION. We have illustrated some of our recent work on statistical methods for subjective testing and summarized and referred to new developments by other authors. We believe that our discussions indicate the direction of research and thinking on problems in subjective testing. We have made one notable omission at least in referring to current research in this area and now note the work of Ferris (1956). Ferris comments in detail on problems in subjective testing and reviews much of the literature of importance. His contributions deal with the construction and analysis of statistical designs in the field of taste-testing. In the abstract of his thesis, he notes

> "Three models of the analysis of variance are put forward as appropriate, illustrating respectively
> (i) how classical latin square and incomplete block designs may be modified to incorporate feature (f) above [the phenomenon to carry-over or residual effects ("after-taste")], recommended especially when various food-samples are being tasted serially for flavor;
> (ii) how the feature (e) [the psychological phenomenon of adaption] may be incorporated in judging various samples of food set out simultaneously, for color, viscosity, or other visually determinable physical characteristic;
> (iii) how one may find a suitable design even when physical considerations impose severe limitations on the choice of statistical designs, as in the case of incomplete block designs of two limits".

We have further research in progress. Still is considering the correlation between the Fisher and Yates' scores for ranks and ranks and between the scores and variate values for various populations. Stuart (1954) had previously considered the correlation between variate values and ranks. Is is possible that this study may yield additional light on the use of the transform of ranks to scores.

Pendergrass has considered the use of discrete scoring scales and the possible loss in efficiency in using scores instead of actual observations on a continuous variable, on the assumption that such observations could have been available. He is also working on the

extensions of the model for paired comparisons to ranking in triple comparisons or in incomplete blocks of size three. In terms of parameters and notation similar to those discussed in the section on paired comparisons, the appropriate model for triple ranking seems to yield the probability,

$$P(X_i < X_j < X_k) = \frac{\pi_i^2 \pi_j}{(\pi_i^2\pi_j + \pi_i^2\pi_k + \pi_j^2\pi_i + \pi_j^2\pi_k + \pi_k^2\pi_i + \pi_k^2\pi_j)}.$$

While it appears that the mathematics associated with this model may be developed, it remains to be seen whether or not application will be simple enough for applied use.

7. ACKNOWLEDGEMENTS. T. S. Russell and C. Y. Kramer generously contributed their time to the preparation of the numerical examples in sections 3 and 4 of this paper respectively. We greatly appreciate this assistance.

REFERENCES

Abelson, R. M., and Bradley, R. A. (1954). "A 2 x 2 Factorial with Paired Comparisons," Biometrics 10, 487-502.

Bose, R. C., Clatworthy, W. H., and Shrikhande, S. S. (1954). "Tables of Partially Balanced Designs with Two Associate Classes," N. C. Agr. Exp. Stat. Tech. Bull. No. 107.

Bradley, R. A. (1953). "Some Statistical Methods in Taste Testing and Quality Evaluation," Biometrics 9, 22-38.

_____(1954). "The Rank Analysis of Incomplete Block Designs. II. Additional Tables for the Method of Paired Comparisons," Biometrika 41, 502-537.

_____(1954a). "Incomplete Block Rank Analysis: On the Appropriateness of the Model for a Method of Paired Comparisons," Biometrics 10, 375-390.

_____(1955). "The Rank Analysis of Incomplete Block Designs. III. Some Large-Sample Results on Estimation and Power for a Method of Paired Comparisons," Biometrika 42, 450-470.

_____(1955a). "Some Notes on the Theory and Application of Rank Order Statistics," Part I, Ind. Qual. Control 11, 12-16; Part II, Ind. Qual Control 11, 5-9.

Bradley, R. A. and Schumann, D. E. W. (1956). "The Comparison of the Sensitivities of Similar Experiments: Applications," To be published in Biometrics.

Bradley, R. A. and Terry, M. E. (1952). "The Rank Analysis of Incomplete Block Designs. I. The Method of Paired Comparisons," Biometrika 39, 324-345.

Calvin, L. D. (1954). "Doubly Balanced Incomplete Block Designs for Experiments in which the Treatment Effects are Correlated," Biometrics 10, 61-88.

Cochran, W. G. (1943). "The Comparison of Different Scales of Measurement for Experimental Results," Annals of Math. Stat. 14, 205-216.

Durbin, J. (1951). "Incomplete Blocks in Ranking Experiments," Br. J. of Psych. (Stat. Sec.) 4, 85-90.

Dykstra, O. (1956). "A Note on the Rank Analysis of Incomplete Block Designs - Applications beyond the Scope of Existing Tables," Biometrics 12, 301-306.

Ehrenberg, A. S. C. (1950). "The Unbiased Estimation of Heterogeneous Error Variances," Biometrika 37, 347-357.

Ferris, G. E. (1956). "Statistical Designs in Taste Testing," Ph.D. Dissertation, Cornell University.

Fisher, R. A. and Yates, F. (1948). Statistical Tables for Biological, Agricultural, and Medical Research, Oliver and Boyd, Edinburgh.

Grubbs, F. E. (1948). "On Estimating Precision of Measuring Instruments and Product Variability," J. of Amer. Stat. Assoc. 43, 243-264.

Hopkins, J. W. (1950). "A Procedure for Quantifying Subjective Appraisals of Odor, Flavor, and Texture of Foodstuffs," Biometrics 6, 1-16.

_____(1954). "Incomplete Block Rank Analysis: Some Taste Test Results," Biometrics 10, 391-399.

Hopkins, J. W. and Gridgeman, N. T. (1955). "Comparative Sensitivity of Pair and Triad Flavor Intensity Difference Tests," Biometrics 11, 63-68.

Kauman, W. E., Gottstein, J. W., and Lantican, D. (1956). "Quality Evaluation by Numerical and Subjective Methods with Application to Dried Veneer," Biometrics 12, 127, 153.

Kendall, M. G. (1948). Rank Correlation Methods, C. Griffin and Co., London.

Kramer, C. Y. (1955). "A Method of Choosing Judges for a Sensory Experiment," Food Research 20, 492-496.

_____(1956). "Additional Tables for a Method of Choosing Judges for a Sensory Experiment," Food Research 21, 598-600.

Kramer, C. Y. and Bradley, R. A. (1956). "Examples of Intra-Block Analysis for Factorials in Two-Associate Class Group Divisible Designs," Submitted for publication.

_____(1956a). "Intra-Block Analysis for Factorials in Two-Associate Class Group Divisible Designs," To be published in Annals of Math. Stat.

Russell, T. S. and Bradley, R. A. (1956). "One-Way Variances in a Two-Way Classification," Submitted for publication.

Scheffé, H. (1952). "An Analysis of Variance for Paired Comparisons," J. of Amer. Stat. Assoc. 47, 381-400.

Schumann, D. E. W. and Bradley, R. A. (1956). "The Comparison of the Sensitivities of Similar Experiments: Theory," Submitted for publication.

Stuart, A. (1954). "The Correlation between Variate-Values and Ranks in Samples from a Continuous Distribution," Br. J. of Psych. (Stat. Sec.) 7, 37-44.

Terry, M. E., Bradley, R. A., and Davis, L. L. (1952). "New Designs and Techniques for Organoleptic Testing," Food Tech. 6, 250-254.

Wilkinson, J. W. (1956). "Analysis of Paired Comparison Designs with Incomplete Repetitions," Ph.D. Dissertation, Univ. of North Carolina.

APPLICATIONS OF ORDER STATISTICS IN MEDICAL EXPERIMENTS

B. G. Greenberg and A. E. Sarhan
Department of Biostatistics
University of North Carolina

1. Introduction. The use of the term "order statistics" connotes various meanings and is here defined to assure understanding in its present usage.

Order statistics is that body of knowledge utilizing the rank or order of an observation as well as its magnitude. In other words, it is a combination of the techniques used in conventional statistics (which consider the magnitude of the observation) with those of rank order statistics (which consider only the relative rank of the observation).

A detailed bibliography of contributions to order statistics is not presented here but several lists may be found in Mosteller [13], David and Johnson [6], and a recent doctoral dissertation by Lott [11].

2. Objective. The purpose of this paper is to illustrate for the applied statistician how to employ recent developments in order statistics for typical statistical problems.

The first two examples will be selected to illustrate how order statistics can be a powerful tool when observations are censored; that is, the exact value of some observations are unknown because a barrier has been imposed by the observer or the measuring process.

The third example will be chosen to illustrate how the use of order statistics can enable the experimenter to estimate the mean and standard deviation of a distribution with high efficiency without the tedium of using all observations in a sample.

The last illustration will be an application of order statistics to the problem of grouping observations into a frequency distribution.

The application of these techniques will arbitarily be restricted to the normal and single exponential distributions since probably they are of most practical value. Order statistics have been applied, nevertheless, to other distributions, (e.g. Sarhan [18], [19]), and the problem of truncation and censoring has been considered in other distributions such as the chi-distribution in Cohen [3], the Type III in Cohen [1], the Poisson in Raj [16] and Cohen [2], and response-time distributions in Sampford [17].

3. Censored Observations. The word censored is applied to instances where sampling is from an unrestricted population, but the exact magnitude of specific observations in the sample may be unknown. The number of censored observations in the sample is known and their ranking relative to some point of censorship is also available.

Censored is different from the term truncated which is applied to instances where sampling is from a restricted population so that the exact numbers that would have occurred in the sample above and below the truncation point are not known. Censored was first used in this context by Hald [9] at the suggestion of Kerrich.

This difference between censoring and truncated is best emphasized by an illustration. If one were to measure the heights of American military personnel, the sample would be from a truncated population of heights because members of that group have qualified by falling between certain minimum and maximum allowable heights. In an industrial context, if one were to select samples from lots that had undergone quality control checks to assure that the manufactured units fell within specifications, the sample would again be from a truncated population, other than for those samples accepted but which should have been rejected.

In measuring the incubation period of a disease, or in life-testing, the experimenter may not have sufficient time or facilities to await the development of the phenomenon in all cases, and might censor the observations on the d^{th} day (Type I) or after p per cent of the observations had responded (Type II). Censoring is usually practiced at the extremes of the distribution. This illustration of censoring observations in life-testing situations is with the exponential distribution.

Censoring with the normal distribution might be for the same reasons or for others as equally important. In certain industrial applications, (e.g. tensile strength) the cost of measuring an observation at the extreme of the distribution is relatively higher than elsewhere. That is, the cost of an observation might be functionally related to its distance from central tendency, and extreme observations are uneconomical to justify. Another reason for censoring a normal distribution might be termed "precision censoring." The measurement error at the extremes of a normally distributed variable might be considerably greater than that of the observations toward the center by having a flat-U-shaped distribution. This occurs in some situations where measurement of physiological functions of the body are involved. Counterparts for precision censoring in industrial and other applications undoubtedly are also available.

The first example in censoring is with the single exponential distribution, and the data are taken from Sarhan and Greenberg [21]. The number of days incubation period resulting from an inoculation is a measure of the amount and potency of the inoculum as well as the individual susceptibility of the test animal. Below are listed the unordered responses from ten rabbits inoculated with a solution containing (0.2) 10^3 treponema pallidum:

Rabbit No.	Incubation in days
1	< 18
2	18
3	>45
4	< 18
5	25
6	21
7	18
8	25
9	25
10	21

Estimates of the earliest possible incubation period (α), the mean ($\alpha + \sigma$), and the standard deviation (σ) are desired from the two-parameter single exponential distribution having the following function:

$$f(y) = \begin{cases} \frac{1}{\sigma} e^{-\frac{y-\alpha}{\sigma}}, & \alpha \leq y \leq \infty \\ 0, & \text{otherwise} \end{cases}.$$

Coefficients from tables provided in the same paper from which these data came make estimation of the two parameters possible despite the censoring of three observations. The observations are rearranged in size order and the coefficients written alongside as follows:

Ordered observations	α	σ	$(\alpha + \sigma)$
< 18	-	-	-
< 18	-	-	-
18	3007/2160	-7/6	487/2160
18	-121/2160	1/6	239/2160
21	-121/2160	1/6	239/2160
21	-121/2160	1/6	239/2160
25	-121/2160	1/6	239/2160
25	-121/2160	1/6	239/2160
25	-242/2160	2/6	478/2160
>45	-	-	-
Estimate	16.10	5.67	21.76
Variance (in terms of σ^2)	0.0567991	0.1666667	0.1114288
Efficiency relative to complete sample	19.56	66.67	89.74

The calculations are as follows:

$$\alpha^* = \frac{1}{2160}\left[(3007)(18) - 121(18) - 121(21) - \ldots - 242(25)\right] = 16.10$$

$$\sigma^* = \frac{1}{6}\left[-7(18) + 1(18) + 1(21) + \ldots + 2(25)\right] = 5.67$$

$$(\alpha + \sigma)^* = \frac{1}{2160}\left[487(18) + 239(18) + 239(21) + \ldots + 478(25)\right] = 21.76$$

The variances of each estimate are found in the tables of the same paper and are also shown in the above table in terms of σ^2. Below the variances of each estimate, the efficiency relative to the complete uncensored sample is indicated. For example, the estimation of α is only 19.56 percent efficient

because two observations were missing on the left and one on the right. The most valuable observations in estimating the start of the distribution, viz. α, should be expected to occur at the left hand side of the distribution. This is verified by the fact if all three missing observations had occurred on the right hand side of the distribution instead of two on the left and one on the right, the efficiency relative to the complete sample would rise from 19.56 to 95.24 percent.

The censoring in this example was of the Type I variety, i.e. employing fixed points on the abscissa. The coefficients used to estimate the parameters, however, were based upon the assumption of Type II censoring. This raises an important question whether a possible bias exists and how much lower the precision is because the known points of truncation, viz. 18 and 45 days, are not utilized in the estimating process.

Several authors, e.g. Sampford [17], have expressed the opinion that the difference between the two is of no great import. The exact solution of the loss in precision is a difficult problem, and work is in progress to measure it. In the interim, we have conducted a sampling study to investigate whether there is a bias, we well as the magnitude of the imprecision. As a result of this investigation, we feel that there is no bias and the order of magnitude of the imprecision is small, particularly in large samples.

The sampling study consisted of 40 samples of size 10, selected from Rand's Table of 100,000 Normal Deviates ($m = 0$, $\sigma = 1$), and estimates of the mean and standard deviation by both Type I and Type II censoring were compared in each sample. For instance, below is a segment, chosen at random, from the sampling study.

Sample No. 36	Censored* at - 0.253 and ordered
0.088	(-1.729)
-0.331	(-1.075)
-1.729	(-0.467)
1.209	(-0.331)
0.840	-0.118
-1.075	0.082
0.894	0.088
-0.118	0.840
0.082	0.894
-0.467	1.209

* The ( ) indicates that the value was censored.

A comparison of the estimates of the mean and standard deviation calculated from each sample included the following: The uncensored data; maximum likelihood method of Cohen [4] for Type I censoring; the method of Ipsen [10] for Type I censoring; and the best linear estimate (BLE) with minimum variance for Type II censoring.

The tabulation below gives some idea of the comparisons for the given sample No. 36, thus indicating why it is thought that there is no bias.

Method of estimation	Mean	Standard deviation
Population	0	1.0000
Uncensored sample	-0.0607	0.9520
Censored		
Cohen (Type I)	-0.0283	0.8042
Ipsen (Type II)	-0.0362	0.8375
B.L.E. (Type II)	-0.0099	0.8527

Calculation of the BLE of the mean and standard deviation employed in the foregoing sample for the normal distribution can best be demonstrated by application to another example. The data are taken from Sarhan and Greenberg [20] and represent Type I censoring at both extremes of the sample.

The observations consist of ten individual systolic blood pressures which were performed by persons learning to measure such readings. Owing to the relatively larger measurement error known to exist for beginners at the extremes, the data thought to be less than 105 mm. and greater than 125 mm. were censored. This resulted in censoring two observations on the left and three on the right.

The data have been arranged in size order, and alongside of them are the coefficients to estimate the mean and standard deviation as follows:

Ordered observations	μ	σ
1. -	0	0
2. -	0	0
3. 108	.20496319	-.88982266
4. 111	.10382533	-.11005067
5. 119	.11220127	-.02620385
6. 121	.11982080	.05494874
7. 125	.45918942	.97112842
8. -	0	0
9. -	0	0
10. -	0	0
Estimates	118.9	16.61
Variance (in terms of σ^2)	.11795477	.17132071
Efficiency relative to complete sample	84.78	33.62

From the percentage efficiency given in the table, the estimate of the mean was 84.78 percent relative to the complete sample despite the omission of 50 percent of the observations. The estimate of the standard deviation does not fare as well, for its efficiency drops to 33.62 percent.

4. Simplified Statistics. In the foregoing paragraph, mention was made of the fact that the estimate of the mean from the sample was relatively efficient although only 50 percent of the sample was used. The optimum combination of half of the sample elements, if the estimation of the mean were to be maximally efficient, might not be the five observations actually used. In fact, owing to the impetus given by Mosteller [13], a whole branch of linear systematic statistics has recently developed in which the purpose is to make efficient estimates of the mean and standard deviation using 2, 3, 4, ..., $k < n$ sample elements.

The most readily identified measure of linear systematic statistics is, of course, the median as an estimate of location. An estimate of dispersion might be the range, semi-interquartile distance, and others. We shall explore these now in a little more detail using data from the exponential distribution as an illustration.

The data come from Table I of Maguire, Pearson and Wynn [12], and represent the time intervals in days between explosions in mines involving more than 10 men killed from December 6, 1875 to May 29, 1951. The 109 observations have been rearranged in size order as follows:

Table 1. Time interval in days between explosions in mines involving more than 10 men killed

Order	Observation	Order	Observation	Order	Observation	Order	Observation
1	1	31	59	61	188	91	354
2	4	32	59	62	189	92	361
3	4	33	61	63	195	93	364
4	7	34	61	64	203	94	369
5	11	35	66	65	208	95	378
6	13	36	72	66	215	96	390
7	15	37	72	67	217	97	457
8	15	38	75	68	217	98	467
9	17	39	78	69	217	99	498
10	18	40	78	70	224	100	517
11	19	41	81	71	228	101	566
12	19	42	93	72	233	102	644
13	20	43	96	73	255	103	745
14	20	44	99	74	271	104	871
15	22	45	108	75	275	105	1205
16	23	46	113	76	275	106	1312
17	28	47	114	77	275	107	1357
18	29	48	120	78	286	108	1613
19	31	49	120	79	291	109	1630
20	32	50	123	80	312		
21	36	51	124	81	312		
22	37	52	129	82	315		
23	47	53	131	83	326		
24	48	54	137	84	326		
25	49	55	145	85	329		
26	50	56	151	86	330		
27	54	57	156	87	336		
28	54	58	171	88	338		
29	55	59	176	89	345		
30	58	60	182	90	348		

The single one-parameter exponential distribution represented by

$$f(x) = \frac{1}{\sigma} e^{-\frac{x}{\sigma}} \qquad x > 0$$

$$0 \qquad \text{, otherwise}$$

has been shown to fit these data quite nicely. If the two-parameter exponential distribution must be used, a simple transformation of location can be used.

The estimate of the standard deviation σ, using all observations, is equal to 241 days. To estimate the value of σ using $k < n$, tables for $k = 1, 2, \ldots, 15$ are available for exponential distribution in a recent paper by Ogawa [15]. For example, if $k = 5$ were selected, the relative efficiency to the complete sample estimate would be 94.76 percent. From Ogawa's table of optimum spacings for $k = 5$, one also learns that the five sample observations which are best to use are as follows:

$$k_1 = (109)(.39347) + 1 = 43$$

$$k_2 = (109)(.67044) + 1 = 74$$

$$k_3 = (109)(.84433) + 1 = 93$$

$$k_4 = (109)(.94387) + 1 = 103$$

$$k_5 = (109)(.98855) + 1 = 108$$

The coefficients in the above formation, viz. .39347, .67044, ..., .98855 were obtained from Ogawa's Table II and the calculations are rounded to the lowest whole integer. Using that same table for the weighting coefficients, the calculations for σ^* are as follows:

Observation Number	Observation (X_{n_i})	Coefficient (a_i)
43	96	.33051
74	271	.21896
93	364	.13173
103	745	.06668
108	1613	.02286

Then, $$\sigma^* = \frac{1}{0.9476} \sum_{i=1}^{5} (X_{n_i})(a_i)$$

$$= \frac{1}{0.9476} (225.56662) = 238.0$$

This compares favorably with the value of 241 calculated from the complete sample. If $k = 10$ were chosen, the efficiency would have risen to 98.32 percent and the value of $\sigma^* = 242.6$.

If the parameters of the normal distribution are to be estimated by simplified statistics, an earlier paper by Ogawa [14] provides the optimum spacings for that distribution. Although Ogawa's spacings are optimal, other combinations of sample observations may be much more convenient to use without any great sacrifice in precision. Such systematic statistics can be found in Mosteller [13], Dixon [7], and Lott [11].

5. Grouping. The optimum spacings used in the previous section for the best linear estimate of the parameters have been shown recently to have another most interesting property in application to a normal distribution. Suppose there are available observations on a continuous variable, and it is desired to classify them, or the population from which they were drawn, into k groups. This might be done either for purposes of convenience in exposition, calculation, or for simplification in the collection of further observations. Thus, if heights of individuals are to be classified as tall, medium and short, the problem is to locate the dividing lines to be drawn without restricting ourselves either to equally-spaced groups or groups with equal expectation of frequency of observations. The criterion for grouping is that if the observations in a group are to be represented by a group central value, the loss in efficiency by this procedure should be a minimum. Furthermore, if this classification is carried out, the loss of efficiency in k = 2, 3, 4, ... groups is also of interest.

This same problem occurs when there are measurements on two continuous variables for a given sample. The experimenter, instead of testing the correlation between the two variables, may prefer for reasons of exposition to group the x variable into k classifications so as to maximize the test of the differences in the y variable among groups by an analysis of variance. Regardless of the correlation between the two variables, D. R. Cox [5] has recently shown that the solution for grouping the x variable comes down to the problem of optimum spacings if the distribution is normal. Thus, if sample data were available on heights and weights, classification of the individuals into tall, medium, and short could be accomplished optimally by dividing the range of heights such that the tall and short groups each had 27.027 percent of the observations and the medium group had the remaining 45.946 percent. This means that the limits of the three intervals would be as follows in a unit normal distribution:

Short: $-\infty$ to -0.612

Medium: -0.612 to $+0.612$

Tall: $+0.612$ to $+\infty$

The loss of precision of this arrangement by substituting one group value for the individual observation results in an efficiency of 80.98 percent. In fact, the optimum groups and efficiencies can be obtained for k = 2, 3, ..., 6 in Cox's paper and for k = 2, ... 11 in Ogawa's paper. The former's values are more nearly precise in the last decimal place than those of Ogawa, and certain results of Cox are reproduced in Table 2 here for convenience.

Table 2. Optimum grouping intervals for unit normal distribution with percentage distribution of observations and efficiency

k	Group limits	Percentage distribution	Percentage efficiency relative to exact observation
2	$-\infty$ to 0	.500	63.66
	0 to $+\infty$	.500	
3	$-\infty$ to -0.612	.270	80.98
	-0.612 to $+0.612$	.459	
	$+0.612$ to $+\infty$	.270	
4	$-\infty$ to -0.980	.164	88.25
	-0.980 to 0	.336	
	0 to $+0.980$	.336	
	$+0.980$ to $+\infty$	.164	
5	$-\infty$ to -1.230	.109	92.01
	-1.230 to -0.395	.237	
	-0.395 to $+0.395$	.307	
	$+0.395$ to $+1.230$	.237	
	$+1.230$ to $+\infty$	.109	
6	$-\infty$ to -1.449	.074	94.20
	-1.449 to -0.660	.181	
	-0.660 to 0	.245	
	0 to $+0.660$	.245	
	$+0.660$ to $+1.449$	.181	
	$+1.449$ to $+\infty$	.074	

The information in Table 2 indicates that the substitution of a binomial classification ($k = 2$) for a normal variate results in an efficiency of 63.66 percent. This particular figure of efficiency may be helpful in estimating required sample sizes in some experiments where there is no experience with a variable to be measured but there is some information on a binomial plane.

There are two points worth mentioning about the use of results in this section. The first is that solution for optimum groupings is identical to optimum spacings for the estimation problem in the case of the normal distribution. This does not appear to be true in general, however, and the rectangular distribution is a case in point.

Secondly, after grouping has been performed, whether by the methods outlined here or not, the estimation of the mean and standard deviation will be made by referring the observations in a group to some central value of that group. Ordinarily, Sheppard's corrections are applied during this step if the groups are equidistant. These corrections may lead to inconsistent estimates for both the mean and standard deviation even when the grouping is equidistant. Consistent estimates for both the mean and standard deviation can be made by maximum likelihood according to the method outlined by Gjeddeback [8].

6. Summary. The uses of recent contributions to the techniques found in order statistics have been applied in three instances. The first is in the case of censored observations both with the normal and exponential distributions. The second application involves estimation of the parameters of the same distributions using $k < n$ of the sample elements. The final illustration is of an application of grouping continuous data into frequency classifications.

References

1. Cohen, A. C., Jr., "Estimating Parameters of Pearson Type III Populations from Truncated Samples", Journal of the American Statistical Association, Vol. 45, (1950), 411-423.

2. Cohen, A. C., Jr., "Estimation of the Poisson Parameter from Truncated Samples and from Censored Samples", Journal of the American Statistical Association, Vol. 49, (1954), 158-168.

3. Cohen, A. Clifford, Jr., "Maximum Likelihood Estimation of the Dispersion Parameter of a Chi-Distributed Radial Error from Truncated and Censored Samples with Applications to Target Analysis", Journal of the American Statistical Association, Vol. 50, (1950), 1122-1135.

4. Cohen, A. C., Jr., "On Estimating the Mean and Standard Deviation of Truncated Normal Distributions", Journal of the American Statistical Association, Vol. 44, (1949), 518-525.

5. Cox, D. R., "A Note on Grouping", Paper submitted to Journal of the American Statistical Association.

6. David, F. N. and Johnson, N. L., "Some Tests of Significance With Ordered Variables", to be published in the Journal of Royal Statistical Society.

7. Dixon, W. J., "Order Statistics". Mimeographed notes.

8. Gjeddeback, N. F., "Contribution to the Study of Grouped Observations," Skandinavisk Aktuarietidskrift, Vol. 42, (1949), 136-159.

9. Hald, A., "Maximum Likelihood Estimation of the Parameters of Normal Distribution which is Truncated at a Known Point," Scandinavisk Aktuarietidskrift, Vol. 32, (1949), 119-135.

10. Ipsen, J., Jr., "A Practical Method of Estimating the Mean and Standard Deviation of Truncated Normal Distributions," Human Biology, Vol. 21, (1949), 1-16.

11. Lott, Fred W., Jr., "The Use of a Certain Linear Order Statistic, Related to the Mean Difference, as an Unbiased Estimate of the Standard Deviation in Finite and Infinite Populations", (1955) Doctoral dissertation, University of Michigan, Ann Arbor.

12. Maguire, B. A., Pearson, E. S., and Wynn, A. H. A., "The Time Intervals Between Industrial Accidents", Biometrika, Vol. 39, (1952) 168-180.

13. Mosteller, Frederick, "On Some Useful 'inefficient' Statistics", Annals of Math. Stat., Vol. 17, No. 4 (Dec. 1946), 377-408.

14. Ogawa, Junjiro, "Contributions to the Theory of Systematic Statistics, I." Osaka Mathematical Journal, Vol. 3, No. 2 (Dec. 1951), 175-213.

15. Ogawa, Junjiro, "Determination of Optimum Spacings for the Estimation of the Scale Parameter of the Exponential Distribution Based on Sample Quantiles", Paper submitted to Journal of the American Statistical Association.

16. Raj, Des. "Estimation of the Parameters of Type III Populations from Truncated Samples", Journal of the American Statistical Association, Vol. 48, (1953), 336-349.

17. Sampford, M. R., "The Estimation of Response-Time Distributions: III Truncation and Survival", Biometrics, Vol. 10, (1954), 531-561.

18. Sarhan, A. E., Estimation of the Parameters of a Skewed Distribution by Linear Systematic Statistics", Journal of the American Association, Vol. 50, (March 1955), 196-208.

19. Sarhan, A. E., "Estimation of the Mean and Standard Deviation by Order Statistics". Parts I, II, III, Annals of Math. Stat., Vol. 25, (1954), 317-328; Vol. 26 (1955), 505-511; Vol. 26 (1955), 576-592.

20. Sarhan, A. E. and Greenberg, B. G., "Estimation of Location and Scale Parameters by Order Statistics from Singly and Doubly Censored Samples. Part I. The Normal Distribution up to Samples of Size 10". Annals of Math. Stat., Vol. 27 (1956), 427-451.

21. Sarhan, A. E. and Greenberg, B. G., "Tables for Best Linear Estimates by Order Statistics of the Parameters of Single Exponential Distributions from Singly and Doubly Censored Samples", Paper submitted to Journal of the American Statistical Association.

PROBLEMS IN A PARTICULAR MILITARY FIELD EXPERIMENT

Kenneth L. Yudowitch
Operations Research Office

I should like to commence my remarks with the enunciation of the three principles which I propose for guidance in the design of military field experiments, the subject of this conference. The three principles are: (1) the exploitation of ignorance; (2) the agglomeration of imponderables; and (3) the balance of weights. In case these designations are not patently clear, I shall attempt to illustrate each of the three principles.

I.

The first principle (Fig. 1)* is rooted in the general technical ignorance of our customer, the soldier, who might well quote from Sheridan's Rivals: "Egad I think the interpreter is the hardest to be understood of the two!" It is perhaps a horrid admission which should be classified, but probably not one Lt. Col. per Pentagon ring can define a Graeco-Latin Square. To illustrate the point, consider a simple statistical test which one might apply to a soldier. We offer him a bet on the drawing of straws, demonstrating first a sample drawing of ten straws from a population of many hundreds. The soldier is offered a fifty-fifty bet on the selection of a straw of his choice -- either long or short. Let us suppose the demonstration drawing yields six long straws and four short straws. As any of us here could tell the soldier, all that he can reliably say about the probability of picking a long straw is that it is significantly greater than 24 percent. This is the customary acceptable lower 95 percent confidence limit on the probability of picking a long straw. And yet our investigations show that the soldier will accept the bet and select the long straw. What is further more discouraging is that the soldier will take our money on such bets.

It is clear that the customer frequently ignores such refinements as 95 percent confidence limits. How to deal with such a barbarian? -- Search out what his question really is before phrasing the objective of an investigation. Then design a test to answer only those objectives in the same language in which the question was asked. If he is disinterested in the beauties of symmetrically oriented test designs, let us exploit this ignorance (though it pain our sensitivity) and make the crude minimum design required. Insistence upon revision of the customer's question to an extent which eventually restricts our ability to answer his real-world question is delusory of self-indulgence.

II.

The principle of the agglomeration of imponderables (Fig. 2) is perhaps best clarified with an illustration from immediate experience: Some time ago I was asked to design an experiment which entailed 180,224 sets of conditions of measurement. The experiment requested had the objective of measuring the relative hit probabilities of eight types of ammunition. Also indicated was some interest in the particular effects of various related parameters, such as qualifications of the subjects, positions of firing, conditions of illumination, and a mass of variables associated with the targets.

* Figures appear at end of the articles.

These target parameters are listed here (Fig. 3). Although obviously in true context each of these parameters exists in a continuum, the range of variations must here be represented by a restricted few values, so we begin agglomeration of these imponderables by arbitrarily selecting the numbers of values to be used for each parameter. As indicated, the first four are represented by only two values each. The justification for this limitation is that careful selection of two rather extreme values will permit recognition of the existence and general magnitude of effect of variation of any one of the parameters, and probably permit rough interpolation.

An immediate agglomeration is provided by the context for the last two of the half-dozen parameters, when there is a marked interdependence in these two parameters. As both characteristics are presented in any one target, the number of the combinations of the two parameters is limited to the number of targets. A systematic attempt to represent all of the combinations of values indicated for each of the six (now five) parameters, results in 352 combinations of parameters, which in this experiment would mean 352 different targets. It is clear, however, that the first four parameters, as well as the last two are also ultimately represented in each of the individual targets of the system. Our preliminary investigation also revealed considerable interdependence among all of these parameters. It is then perhaps the ultimate application of the principle of agglomeration of imponderables to dump the variations of all six parameters into one presentation of the target system. As the facts of life limited us to 22 targets, the application of the second principle results in a reduction of from six times infinity to 352 to 22 representations of these half-dozen parameters. Finally then, all 22 targets appear in a single sequence which we call a <u>run</u>.

In addition to the target characteristics, we are concerned also with characteristics of the subjects, the environment and the test materiel. Grouped here are these several parameters (Fig. 4). The subjects come to us in four formal qualifications. From the variety of firing positions as above, we selected two; similarly for the illumination. These three parameters then yield a product of 16 combinations. Qualification varieties were simply deleted from the experiment proper, and four special runs programmed for measurement of variation in this parameter.

The handling of the four possible combinations of position and illumination is a nice illustration of a corollary of the second principle. The little block diagram (Fig. 5) represents the four possible combinations (day and night, sitting and standing). If there is a degree of independence between variations in position and variations in illumination, it is quite possible to infer the value for a fourth box of this square array, given three. In this instance we elected to make measurements for both positions in the daytime and for the sitting position at night. We thus obtained two measures of score degradation, one for the shift of position from sitting to standing, and the second for the shift of illumination from day to night. Presumably the score for the unmeasured category (night standing) is deduced by applying both degradation factors multiplicatively to the day-sitting score. Thus we reduce 16 combinations of these three parameters to only three.

We come to the last two factors, the ammunition itself and variation in subjects (Fig. 6). Our experiment has specified eight ammunitions. The number of samples of subjects required depends on the anticipated variation from sample to sample. As a compromise with practicality, four samples were agreed upon. This number seems sufficient to give a fairly reliable indication of the degree of variation among samples: The average is meaningful if the variations are small; and the opportunity for variation is adequate to indicate whether a larger number of samples was required. There is no simple means of agglomeration here, so that our ammunition and population parameters leave us with 32 separate experimental conditions.

Finally, however, a further agglomeration is made by limiting the number of combinations of each of the four samples with each of the three position-illumination categories. In this case, instead of the twelve possible combinations, only eight of the combinations are selected so that our ultimate number of runs is 3x32x8/12 or 64. In addition, the special qualification runs consist of two ammunitions and two population samples, making a grand total of 68 runs (Fig. 7). We have reduced the number of experimental conditions from a grand total of 16x16x32 or 8192 runs by a factor of 120, by application of the second principle.

As a very heavy schedule permitted a maximum of 8 runs per day, the schedule of 68 runs (1496 conditions) required 8 1/2 days in the field, following preparational field work. It is of interest to note that if this same experiment were attempted without application of the second principle, whatever conclusions might have been reached concerning the test materiel would be totally obsolete; as the 8192 runs (180,224 conditions) would take four years of steady work to complete. -- This is based on a five-day week with Christmas, Independence Day, and Armed Forces Day off.

III.

Inherent in the illustration used for the second principle is the application of the third principle of experimental design, the balance of weights (Fig. 8). Quite obviously among the 180, 224 combinations of parameters possible in the illustrative experiment, there are some combinations rather more important than others. It is essential that we consider not only the nicety of design for simplified analysis procedures, but that we consider why we are doing the experiment in the first place. The customer (who is the quite ignorant fellow we spoke of earlier) is unconcerned with statistical niceties. He is however very much concerned in finding certain answers which are vital for his decisions, and somewhat less interested in finding certain other answers which will be of incidental utility in guiding his decisions or activities. Thus, for example, it may be that the customer is very vitally concerned with comparative capabilities of two of the eight types of ammunition under study, and somewhat less concerned with comparisons involving the other six types. It is incumbent upon the experiment designer to recognize this difference in interest, and to respond to it with appropriate distribution for weighting of experimental effort. Any refusal to consider balance of weighting of experimental effort is patently justified, as justification for the performance of the entire experiment springs only from the interest of the customer. Any weighting of subcategories of interest must in any honest experiment be reflected in the experimental effort.

A second, more technical factor also affects appropriate weights of portions of an experiment. For example, note the 8/12's of the possible combinations of population sample with position illumination which were selected. The diagram (Fig. 9) illustrates the eight out of twelve possible combinations selected. It is clear that emphasis has arbitrarily been placed on the day-sitting runs; half each of the day-standing and night-sitting runs having been deleted. The logical attempt to justify such an asymmetrical procedure is as follows: In the first place the reduction from twelve is strongly urged by practical limitations on the total experimental effort. One might, however, expect a more symmetrical or uniform mode of reduction. But a uniform reduction of the experiment threatens that the resultant measurements may border on statistical unreliability, merely because of the small sample size. The selection made obviously permits all four population samples to be used with <u>one</u> of the conditions of firing, providing a reliable measure for <u>that</u> condition. The other two conditions (day-standing and night-sitting) are less reliably measured. This is justified, because it is much more important to determine whether the ammunition differences sought exist under any condition than it is to determine the variations of this difference with the several conditions of firing.

I should like to close my remarks with a carefully considered statement: "My immediate point is that the questions involved can be dissociated from all that is technical in the statistician's craft, and that when so detached are questions only of the right use of human reasoning powers, with which all intelligent people, who hope to be intelligible, are equally concerned, and on which the statistician as such, speaks with no special authority."

V3717

The Exploitation of Ignorance

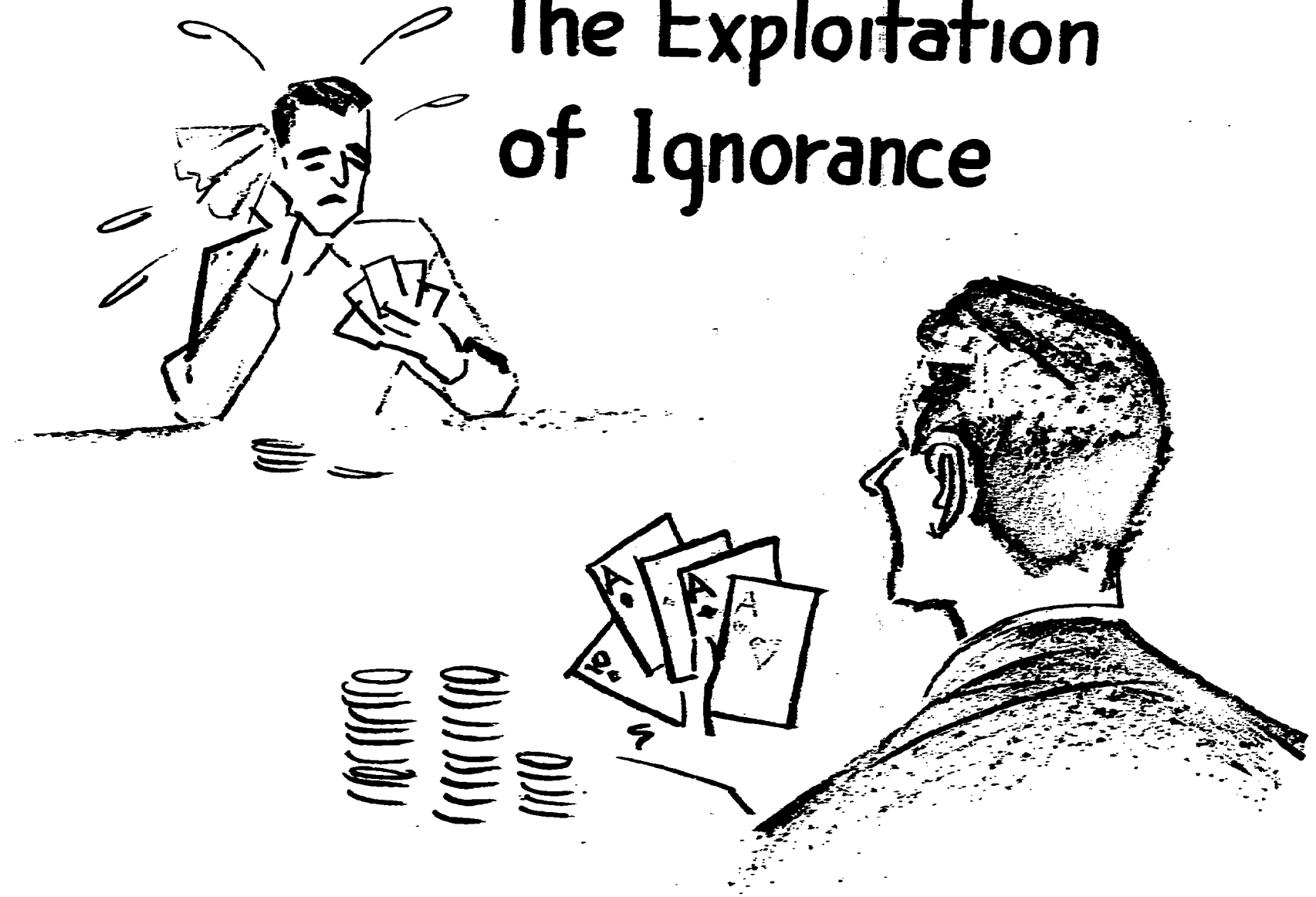


Fig. 1

The Agglomeration of Imponderables

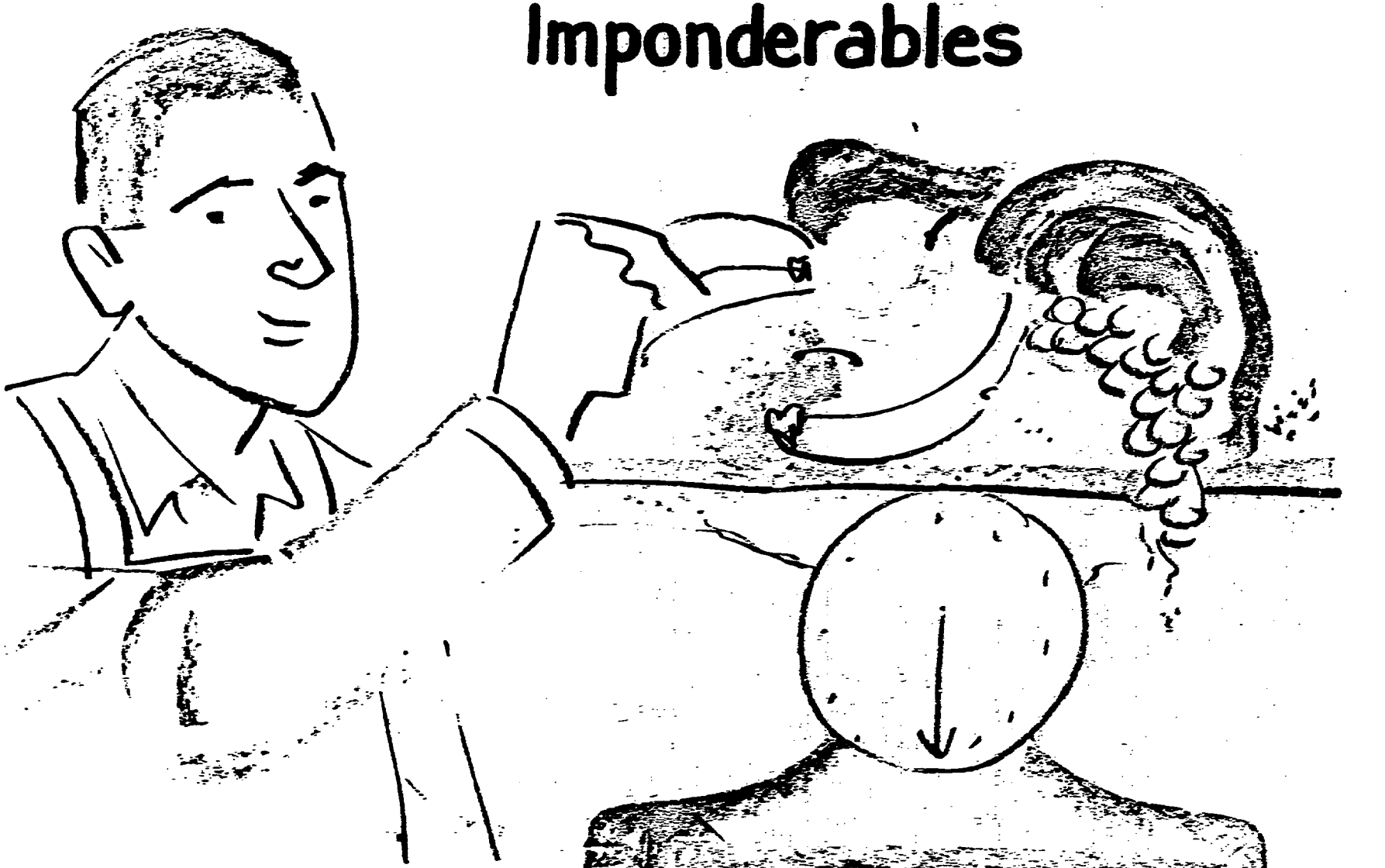


Fig. 2

Target Parameters

1-22 { Size 2
Concealment 2
Movement 2
Activity 2
{ Exposure
Range } 22 } 352-16

Subject and Environment Parameters

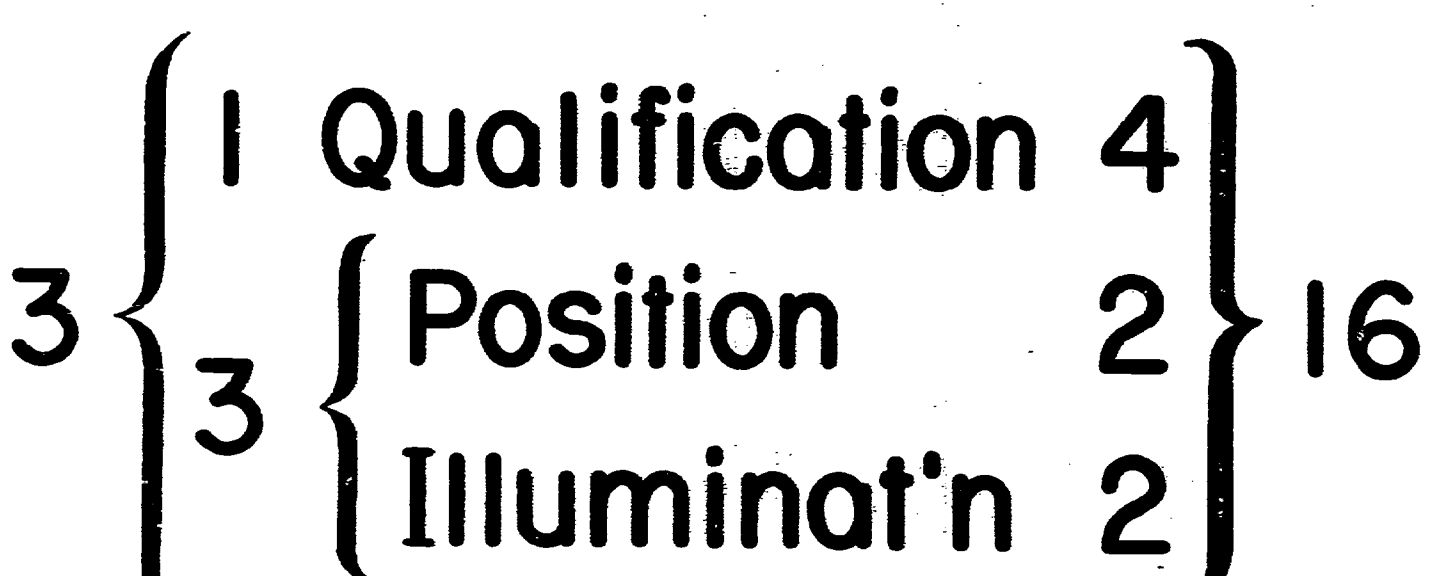


Fig. 4

V3722

FIG. 5

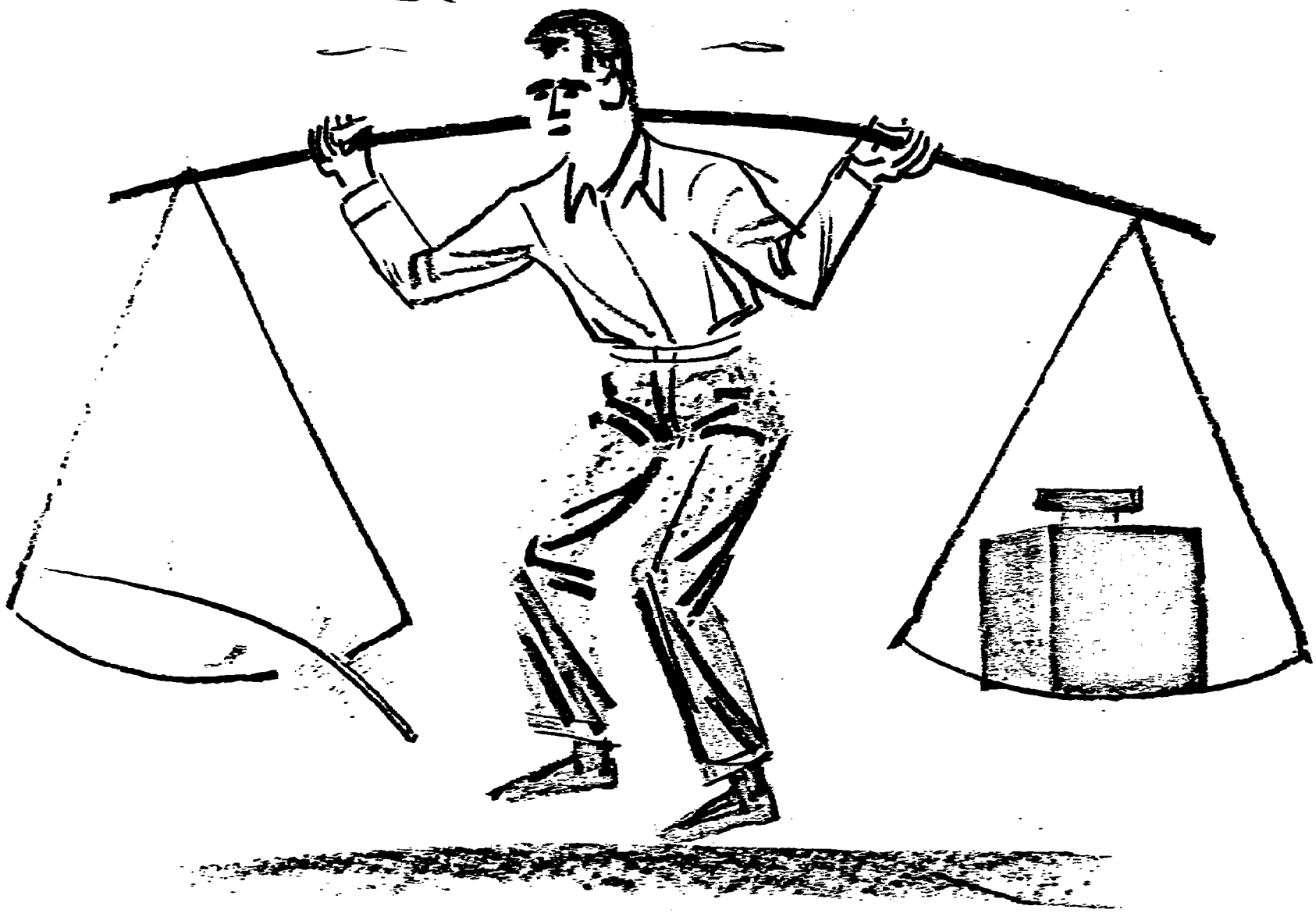
Materiel and Subject Factors

$$32 \left\{ \begin{array}{ll} \text{Ammunition} & 8 \\ \text{Sample} & 4 \end{array} \right\} 32$$

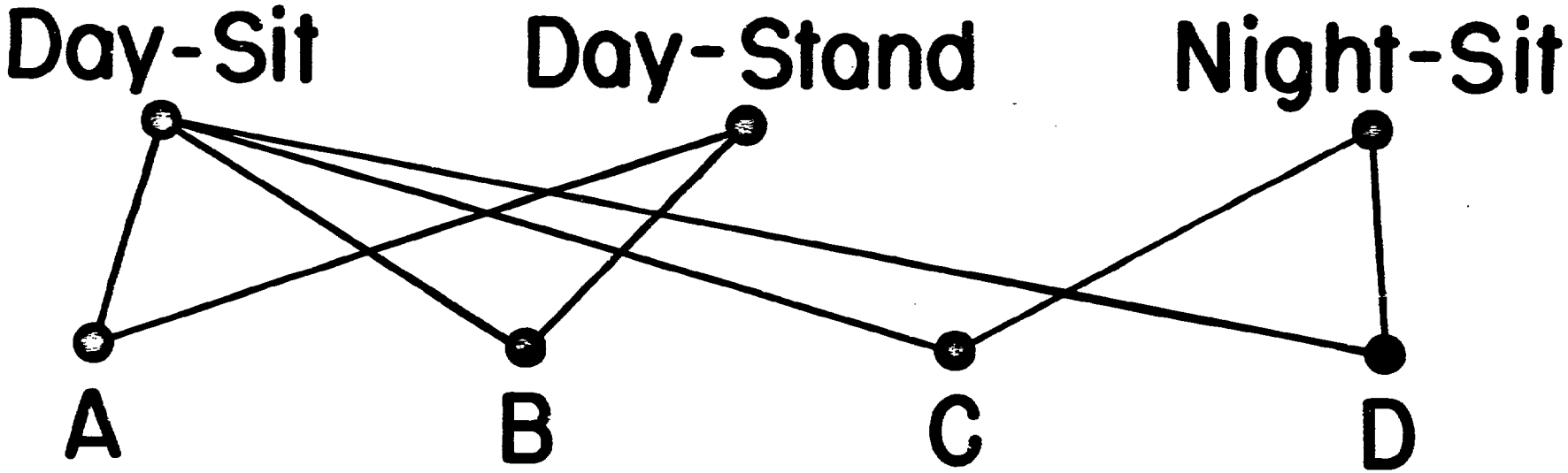
Total Runs

$$16 \times 16 \times 32 = 8192$$

$$1 \times 3 \times 32 \times \frac{8}{12} + 4 = 68$$



The Balance of Weights


Day-Sit
Day-Stand
Night-Sit
A
B
C
D

HUMAN ENGINEERING EXPERIMENT ON TUBE TESTER TV-2/U

Harold Zweigbaum and Donald Donaldson
Signal Corps Engineering Laboratories

The extensive use of electron tubes in military equipments has led to a widespread acceptance of the conventional tube tester as a maintenance tool. In the early days of electronics, when the multi-element electron tube was coming into general use, the adequacy of the tube tester to determine the quality of the tube was quite satisfactory. Operating frequencies were relatively low, and emission or transconductance tests were made under conditions that rather closely approximated the actual operating conditions.

Now, however, not only have the number of tube types increased by several orders of magnitude but the applications have become more complex. Operating conditions vary widely. The use of a conventional tester which allows but one value of plate voltage and two values of screen voltage to be applied to the respective elements has not proven to be of value in military depots. There the need has been for a tube tester that will allow the application of variable voltages to the tube elements, so that a reading of quality can be compared with the manufacturer's original acceptance data.

This requirement was staisfied by the development of the Electron Tube Test Set TV-2/U, featuring continuously variable and metered voltages to the several elements of the tube under test. While the flexibility thus attained permits tube testing within the requirements of MIL-E-1, this desirable operational feature has created human engineering problems.

After the tube tester was built and during its preliminary use at the Signal Corps Engineering Lab, it became apparent that, even though the front panel was laid out in a logical sequence, the number of manipulations required to perform a test on a tube was conducive to error.

In order to ascertain whether or not the operation of the tube tester placed undue reliability on operator capability, a statistical experiment was designed, using the tube tester and actual operators. (See the picture of the front panel layout of the TV-2/U at the end of this paper.)

As we can readily see, the versatility of the TV-2/U was obtained at the expense of an increased number of controls, switches, and monitoring meters. In the more conventional tube tester used for simple maintenance applications, about twenty separate and distinct manual operations are required in order to check the condition of an average receiving tube; this same type of test in the TV-2/U requires that an operator go through about thirty-four separate steps, or about a 75% increase. In order that we might see what effect, if any, in the performance of an operator was due to the increased number of manual operations, an experiment was devised to provide data pertinent to the precision of measurement obtained by a normal class of operators.

In planning the experiment, certain controlling conditions were evident. It was necessary to select individuals whose performance could be considered representative of that group. Further, since individual variation among operators is to be expected, more than one individual was required in order

to permit a measure of the sampling variation. It was also necessary that the equipment under consideration be tested across the range of its intended operation in order to eliminate from the results any spurious homogeneity occasioned by too narrow a range of study. The standard for comparison that was chosen was the set of measurements obtained by laboratory engineers who were familiar with the equipment and its operation.

The specific test schedules involved the choice of twenty-five (25) electron tubes, five (5) from each of five (5) generic groups, pentodes, triodes, voltage regulators, diodes and rectifiers, thus covering most of the operating range for which the TV-2/U is designed. Thyratrons were excluded from the schedules. This group of tubes is one of the most difficult to test, and it was decided to utilize statistical data from other tube types in estimating operator precision. The premise of this decision was that data from the thyratron tube type would be unnecessary should the results from other types prove conclusive.

The test procedure was established in three phases. First, the selected tubes were measured by laboratory engineers (one of the two classes of operators) at Evans Signal Laboratory on two test sets, and the data recorded. Second, the test sets were transported to the Tobyhanna Signal Depot in Pennsylvania. After the initial training of depot personnel (the second of the two classes of operators) in the use of the equipments, several weeks were allowed for familiarization and actual use, during which time these personnel tested over 4000 electron tubes. The sample tubes were then measured on each test set by the depot operators, on separate days and under the observation of a laboratory engineer. Third, after completing and recording these measurements, the tubes and test sets were returned to ESL where the laboratory measurements were repeated and recorded. This latter step insured against tube damage during the testing interval. We should note that at no time were the depot operators aware that they were participating in the experiment. They were merely informed that the tube testers were being checked for ruggedness under constant use.

In order to avoid a possible influence of repetitious measurements on the data, the measurements were performed by both classes of operators on individual tubes, selected in a random fashion.

Analysis of the measurements involved an estimate of the sampling variance for operators within each of the classes. This estimate was to be derived from twenty-five pairs of measurements, the expected values of which were specially chosen to cover the range of use of the tube tester. Each of these estimates of the sampling variance would then be used in an F test to determine whether or not the ratio of estimate values was consistent with an expected ratio derived from two random samples from the same population. The hypothesis tested by the F test is: there is no real difference between the use of the equipment by the laboratory engineers and its use by the depot operators. After the depot data had been recorded, it appeared that the second depot operator possibly had received insufficient training in the use of the tube tester. In approximately twenty-five percent of his trials he was unable to adjust the tube checker so as to obtain a reading. Since the estimate of operator variability is obtained

from paired measurements on the same tube, the data from this twenty-five percent were discarded leaving the data from nineteen tubes available for estimating the variance for depot operators. The corresponding figure for laboratory engineers is 25 tubes.

It is to be noted at this point that this arbitrary deletion of data, deviating grossly from the average, does not of itself invalidate either the results or any conclusions that may be drawn therefrom. In the present experiment it is recognized that the deletion acted to provide an indication of higher reproducibility within the depot operator class than would otherwise be evidenced.

The structure of the experimental design and its analytic procedures have thus been preserved at the cost of reduction of approximately twenty-five percent of the data by limiting the study to those 19 tubes upon which all four operators obtained readings. (See Figures 1 and 2.)

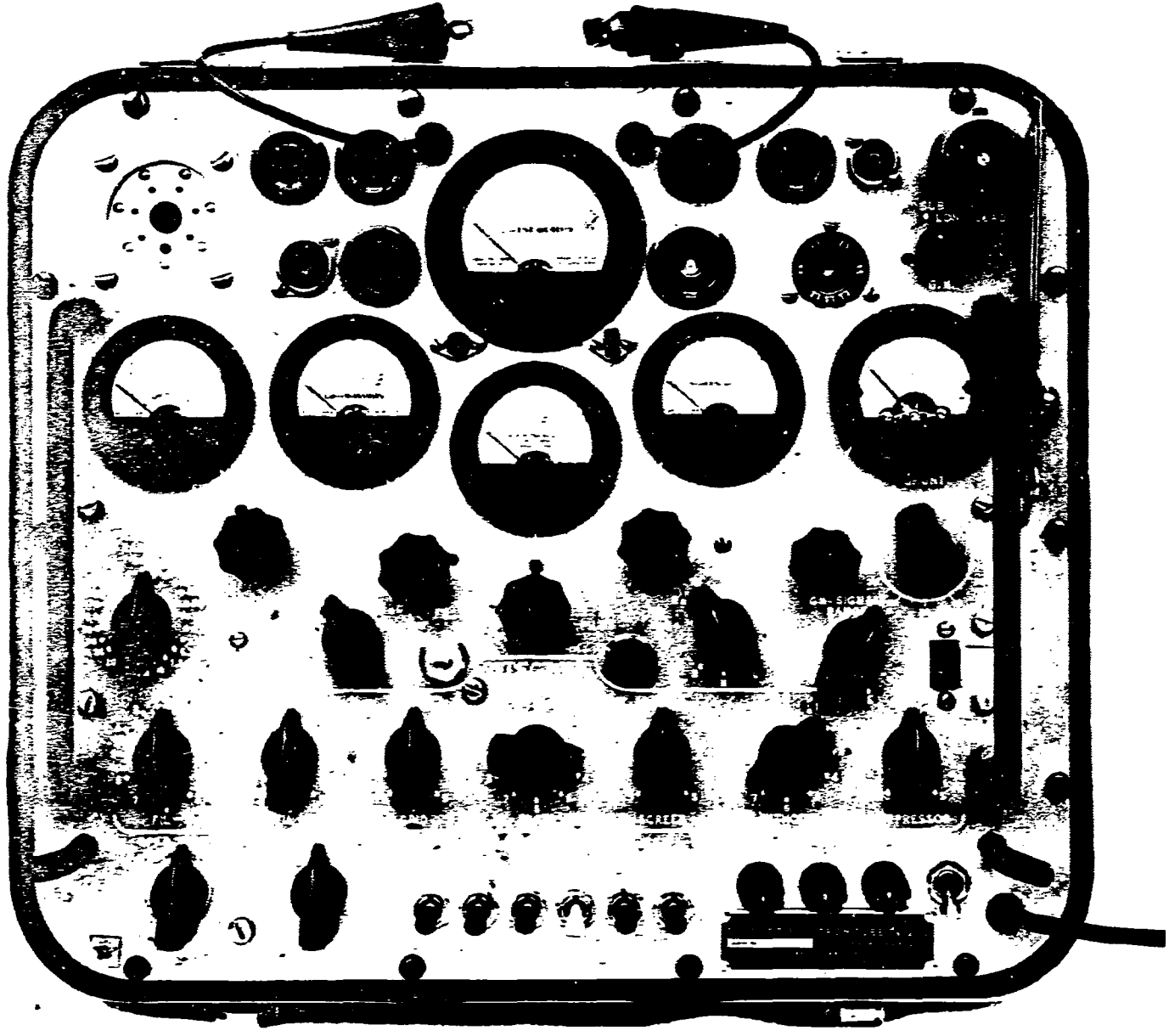
As we can see by the chart, the only major contribution to variation in results, other than that of tube variability which was deliberately introduced, is that variation contributed by the lack of reproducibility within the class of operators.

Finally, an F test was applied to the data in order to confirm or deny the hypothesis of equal precision of classes of operators. (See Components of Variance Table at the end of this paper.)

For eighteen degrees of freedom in each measure of variance, the 5 percent probability, or 95% certainty value, of F is 2.22. The value of F obtained from the measured data is 13.55. This result is definitely significant and denies the tenability of the hypothesis of equal precision of the two classes of operators.

It was concluded from this experiment that the Tube Tester TV-2 is too complex an instrument to be used by depot personnel with any degree of assurance that the results obtained by these operators will accurately reflect the condition of an electron tube.

The results of this experiment show us that in order to have a truly effective tube tester of the TV-2 type, it is necessary to eliminate a high percentage of the operator manipulations. The Signal Corps Engineering Laboratories have instituted a program to study the effects of applying automatic processes to a tube tester. The study will take into account the effects of the human engineering experiment and will be aimed at the practical embodiment of an automatically controlled tube tester that will be at least as small, light, and accurate as the TV-2/U.



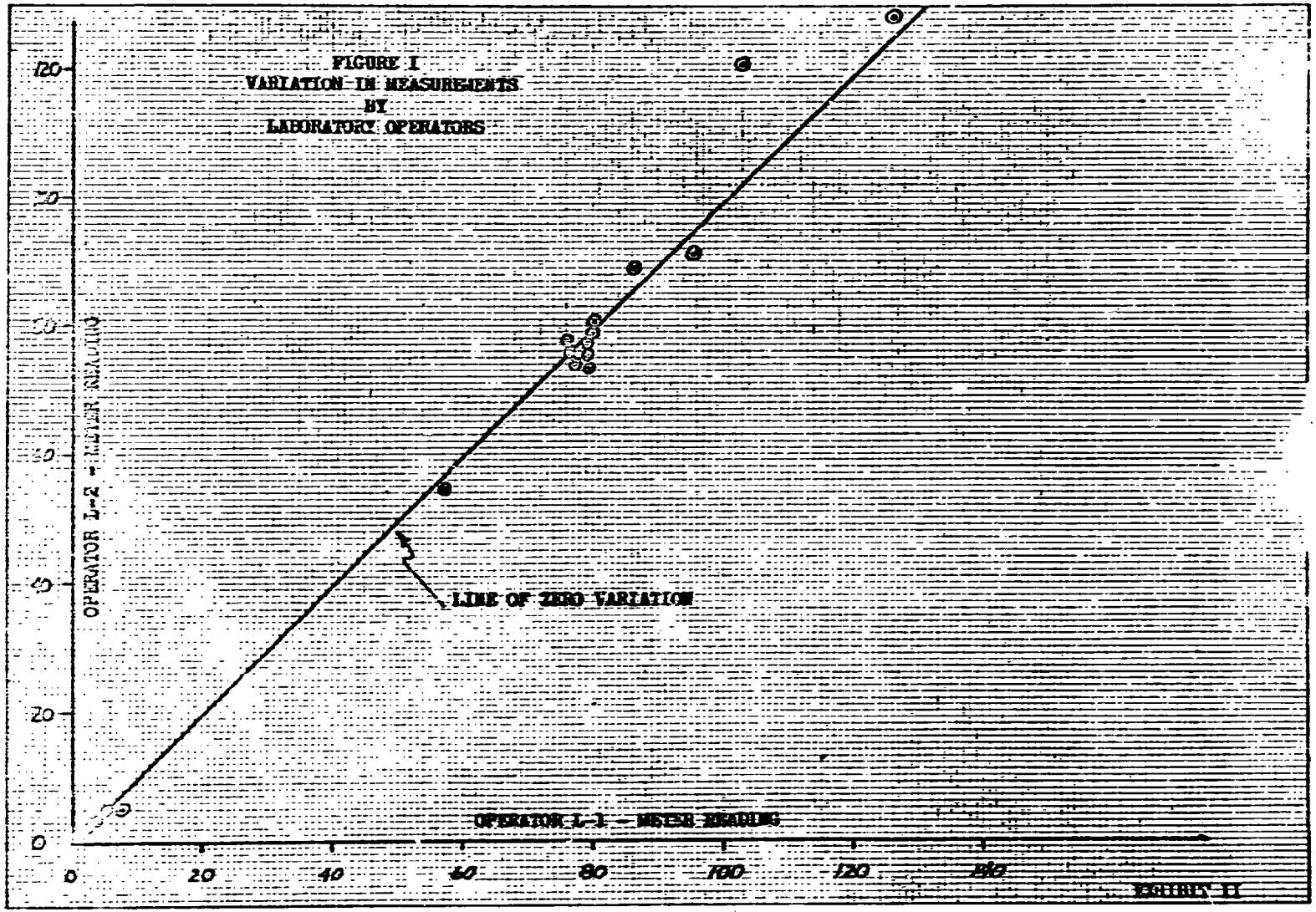


FIGURE I
VARIATION IN MEASUREMENTS
BY
LABORATORY OPERATORS
120
100
80
60
40
20
0
OPERATOR L-2 - METER READING
LINE OF ZERO VARIATION
OPERATOR L-1 - METER READING
0
20
40
60
80
100
120
140
EXHIBIT II

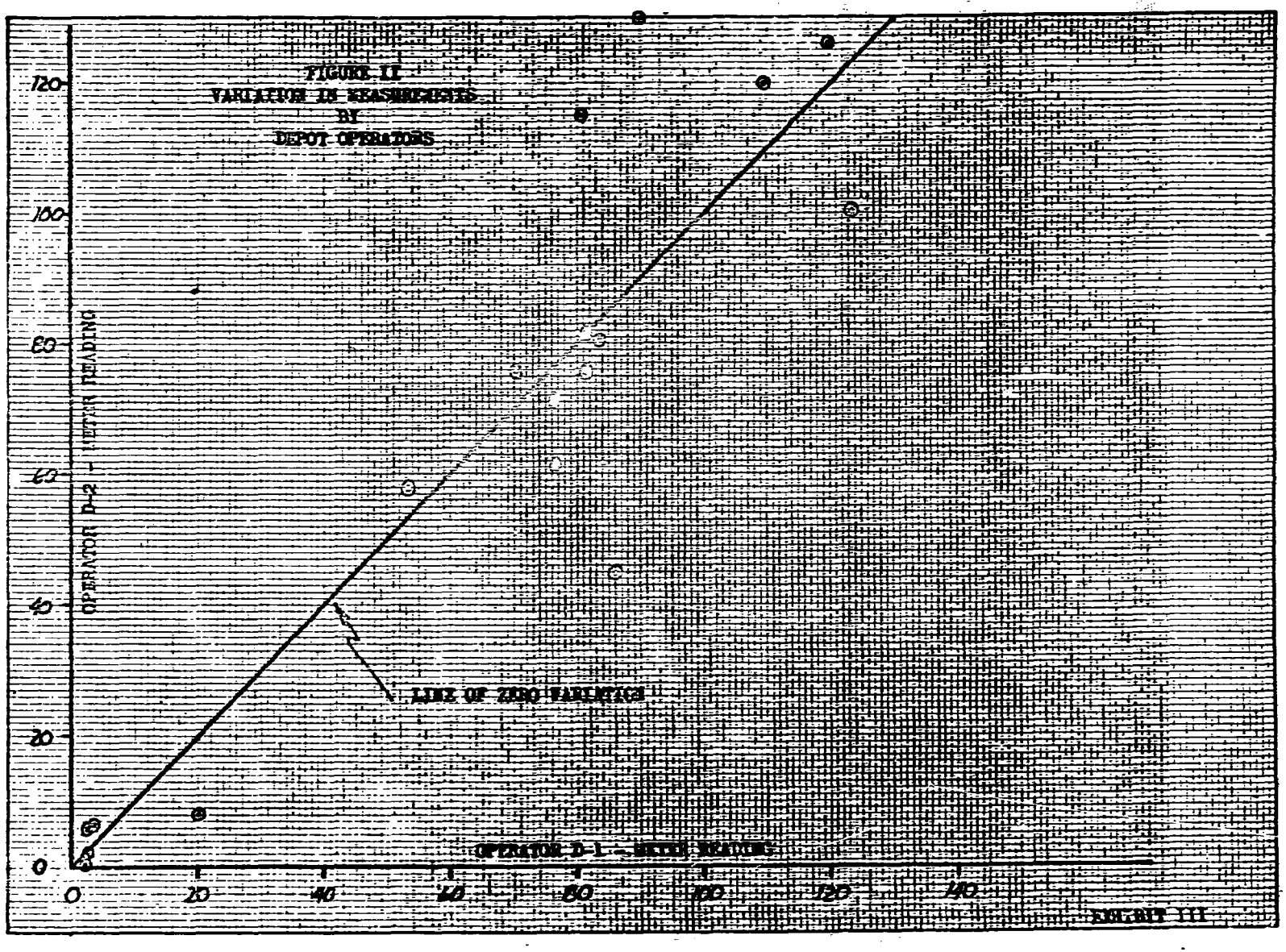


FIGURE II
VARIATION IN MEASUREMENTS
BY
DEPOT OPERATORS

EXHIBIT III

COMPONENTS OF VARIANCE

Components	Sum of Squares	Degrees of Freedom	Mean Squares
Between Classes of Operators	9.24	1	9.24
Operator Sampling	6.47	2	3.285
Within Lab. Operators	5.53	1	5.53
Within Depot Operators	0.94	1	0.94
Tube Variability	119,444.38	36	3,317.899
Within Lab. Operators	57,539.33	18	3,196.663
Within Depot Operators	61,905.05	18	3,439.169
Reproducibility	3,275.15	36	90.076
Within Lab. Operators	224.09	18	12.505
Within Depot Operators	3,051.06	18	169.053
TOTALS	122,735.24	75	

METHOLS OF ESTIMATING LETHAL DOSE FOR MAN

Clifford J. Maloney
Army Chemical Corps

I. INTRODUCTION. It is of interest in medical and biological research to be able to estimate the dose-response curves[1] for the mortality response of humans to various infectious agents. Direct methods of experimentation are neither practical nor ethically permissible, therefore indirect methods of estimation are required. It is the purpose of this paper to show how two routine types of biological measurement can be combined in various ways to produce estimates of human mortality response to infectious agents. The methods have the advantage of being absolute determinations depending on no unverified extrapolations. The types of measurement which are required are: (1) morbidity and mortality rates for man and other animals in natural environments; (2) dose-response experiments for morbidity responses in humans and for morbidity and mortality responses in other animals.

Morbidity statistics can be obtained for many diseases through routine reports to health departments of cases of infectious diseases. Special studies can be conducted to estimate the incidence of poorly reported or non-reportable diseases. Adequate mortality statistics are usually available because death certificates, specifying primary and associated causes of death, are filed for almost one hundred percent of all deaths in the United States. If need be, the death certificate data can be supplemented by special surveys aimed at greater accuracy and/or completeness. The numbers of cases and deaths within a segment of the population readily can be converted to rates on the basis of existing census statistics, on estimates of the current population based on previous census figures, or on special sample surveys or enumerations.

Good experiments involving animal morbidity and mortality can readily be conducted, assuming that the obstacle of funds to purchase and handle quantities of animals is overcome. It is moreover quite possible that animal experiments for other purposes can be exploited. On the other hand, the requirement of fairly large numbers of experimental subjects for the proper determination of dose-response probit lines well may interfere with the conduct of human morbidity experiments. Nevertheless, experiments utilizing modest numbers of human volunteers can be set up, so that desired human morbidity probit lines can be defined, even though their parameters may have more than desirable sampling error.

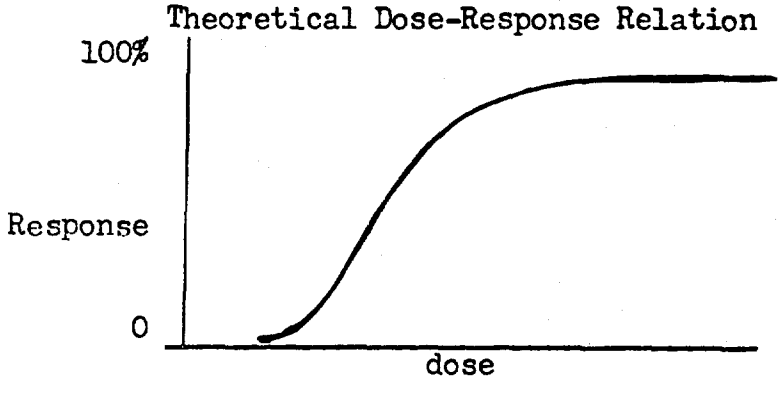
Three suggested methods, based on the types of data described above for estimating human mortality dose-response curves are given below. These methods cannot be expected to produce precise estimates of probit parameters, but they have an advantage over other suggested methods in that they utilize human data to produce estimates of parameters which describe human responses, rather than depending on non-human data for such estimates. One suggested solution to the problem of estimating lethal dose for man requires the

1. See Section II for a discussion of dose-response curves.

assumption that the mortality probit lines for certain simians, or other animals, are "close" to those for humans. This latter method lacks logical justification because we know that there is considerable variation in responses to infectious agents among even closely related animals, i.e., the responses to the same dose of an infectious organism by rhesus monkeys, by cynomologus monkeys, and by chimpanzees may differ markedly from each other. The three methods outlined below are logically unimpeachable.

II. DIGRESSION ON DOSE-RESPONSE RELATIONS. Biological material is notoriously variable. This does not mean however, that it is subject to no law, but only that the law has a statistical character. The response shown by an organism to a hostile influence of physical, chemical, or biological nature is therefore predictable only in terms of mean values over many individual responses. The variability of the response to microbiological agents is greater than that to chemical agents. The average response, of course, increases as the quantity of agent increases. As the response cannot be less than zero nor more than one hundred percent, a plot of the mean dose-response relation would appear as:

Figure 1

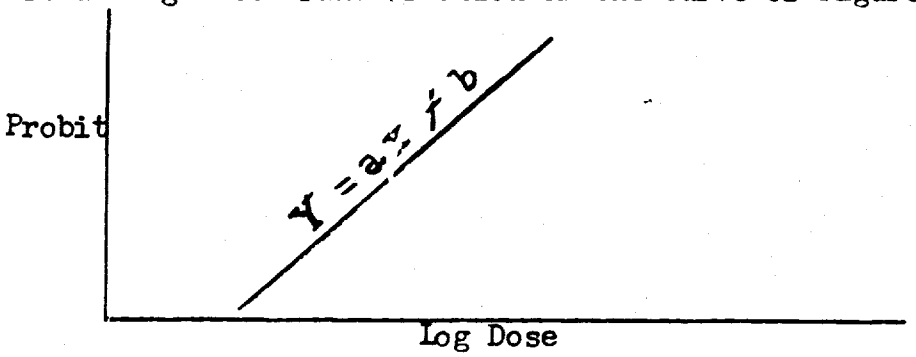


It has been widely verified[2] that converting doses to the corresponding logarithmic values and transforming percent responses by the integral of the normal curve or error usually converts the asymmetric curve of Figure 1 to a straight line.

2. Finney, D. J., "Probit Analysis," 2d edition, Chapter 3.

Figure 2

Probit Log-Dose Transformation of the Curve of Figure 1

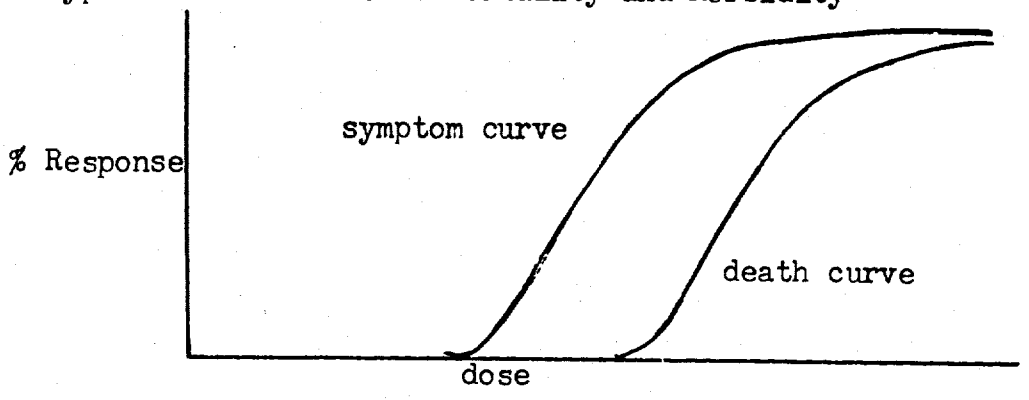


It is of course true of this line, as of any straight line, that it is determined as soon as one point on the line and the slope of the line have been fixed. It is customary to choose for the point the one showing the probit of 50% animal response, since this point requires less experimentation for its measurement to a given degree of accuracy than any other point on the line. This point is known as the 50% endpoint and symbolized as ED50. Infectivity response is then ID50 and mortality response LD50. It is clear that several distinct curves could be plotted on the graph of Figure 1 and that the same transformation would reduce them all to straight lines corresponding to Figure 2.

III. METHOD I. This method arose from the simple consideration that a dose sufficient to kill must be sufficient to provoke symptoms. Hence, if a curve of doses vs. percent showing any chosen symptom syndrome is plotted on the same graph, the two curves cannot cross.

Figure 3

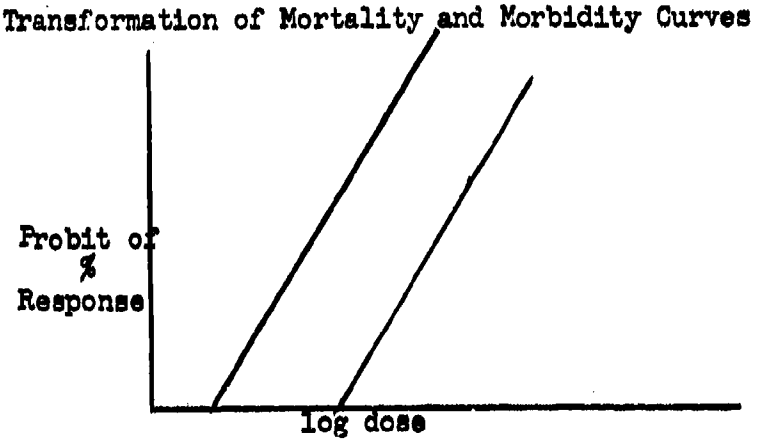
Hypothetical Curves of Mortality and Morbidity



The only point to notice about the graph is that the death curve is beneath the symptom curve at all doses. The probit transformation converts each of these curves to straight lines. Now, as the curves did not intersect, neither will the lines. Hence they are parallel. It is clear that the slope of the mortality line is therefore known because it is the same as that of the morbidity line.

Figure 4

Probit Transformation of Mortality and Morbidity Curves of Figure 3



If the infectivity and mortality probit lines for man for a particular organism are parallel, we can estimate the LD50[3] if we have (1) experimental ID50[4] data and (2) case and death rates observed in nature[5]. The following steps lead to an estimate of the parameters of the mortality probit line:

(a) Fit a probit line ($Y = a x + b$) to the experimental infectivity data.

(b) Compute the case rate in the population (CR).

(c) Compute the death rate in the population (DR).

(d) Using the equation for the infectivity probit line, compute the theoretical infective dose for the observed case rate (IDCR).

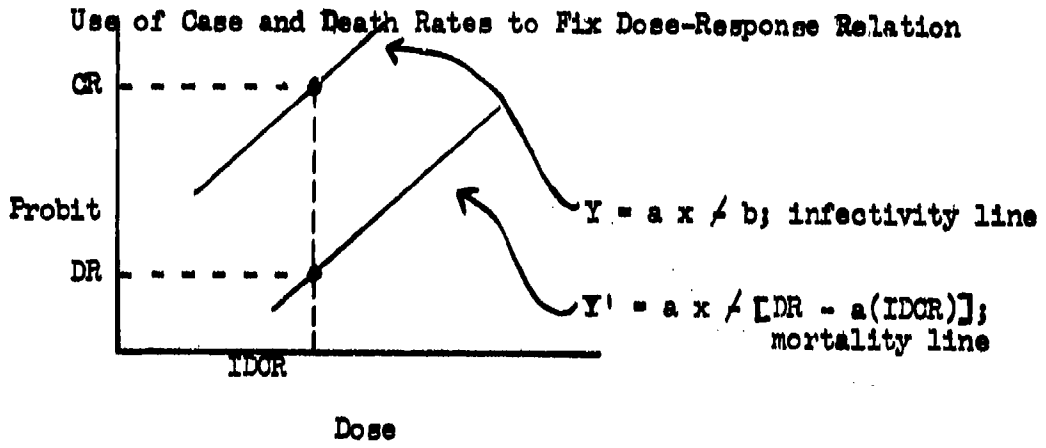
(e) The equation for the mortality probit line can be obtained by using the slope of the infectivity probit line (as the lines are parallel by the theory underlying this procedure) and the intercept is defined by the equation DR = (IDCR) $a + b$, or $b = DR - a(IDCR)$.

3. Dose producing 50 percent deaths.

4. Dose producing 50 percent infection.

5. Using the route of infection which is of interest.

Figure 5



The argument on which parallelism of morbidity and mortality lines is based would not apply in the case of ancillary symptoms which do not lead to death when aggravated. Hence, if such non-parallel lines are found, they might be used to separate the symptom complex into those leading and those not leading to death.

IV. METHOD II. The human dose-response mortality probit line can be estimated by securing a minimum of two measurements of mortality at different measured dose levels. This might be done by measuring exposure and mortality of (1) workers who work in an occupation with a high risk of infection (for example, farmers and ornithosis[6], animal fiber workers and anthrax[7], etc.); (2) laboratory workers exposed to an organism[8]; (3) groups in the normal population exposed to relatively high doses of a causative agent (for example, people living near dairies in Los Angeles and Q fever[9], residents of Leavenworth County, Kansas and histoplasmosis, etc.)[10]. Estimates of the dose to which these people are exposed could be obtained by intensive sampling of their environments. Cause-specific mortality figures could be obtained by routine epidemiological methods.

The observed dosages and death rates would then be used to plot and/or compute a mortality probit line.

---

6. Karrer, H., B. Eddie and R. Schmid, Barnyard fowl as a source of human ornithosis. Case report, Calif. Med., 73(1950):55-57.

7. Dignam, B. S., Anthrax--an industrial disease, Conn. Med. J. 15(1951):316-17.

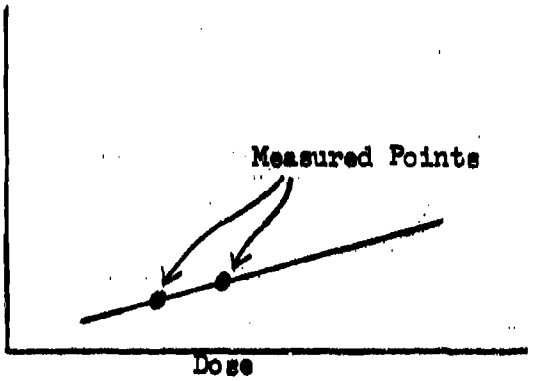
8. Sulkin, S. Edward and Robert M. Pike, Laboratory acquired infections, J.A.M.A., 147(1951):1740-1745.

9. Shepard, Charles C., and Robert J. Huebner, Q fever in Los Angeles County, Am. J. Pub. Health, 38(1948):781-788.

10. Furcolow, Michael L. and Jay Sitterly, Further studies of the geography of histoplasmin in Kansas and Missouri, J. Kansas Med. Soc. (1951).

Figure 6

Hypothetical Determination of Human Dose-Response Curve
Employing Dose Measurements from Natural Environments



It is well to point out that this technique is wholly unrelated to the exploitation of extraordinary laboratory accidents. In fact, such accidents[11] due to recognizable discrete departures from the usual laboratory environment, are to be regarded as unwelcome complicating factors, so far as morbidity and mortality rates are concerned, though contributing fully to the case fatality rate determination. Instead of attempting to infer the dose actually received by cases, the average dose level of exposure both of reactors and of non-reactors would be ascertained by sampling procedures[12]. Then techniques for computing bioassay with error in the dose[13,14] would be used.

V. METHOD III. Method II can be modified so as to eliminate the requirement of direct measurement of dosage in natural environments. This can be done if mortality rates at two unmeasured dose levels are known for both man and for some other animal species, and if a dose-response mortality curve can be obtained experimentally for the same animal species.

The procedure of Method III is as follows:

1) Measure mortality both for humans and for a species of animals at each of two (preferably widely different) dosage levels, say A and B. Call

11. Sabin, A. B. and A. M. Wright, Acute ascending myelitis following monkey bite with isolation of virus capable of reproducing disease, J. Exp. Med. 59(1934):115.

12. Ibach, Martha J., Howard W. Larsh, and Michael L. Furcolow, Epidemic histoplasmosis and airborne Histoplasma capsulatum, Proc. Soc. Exp. Biol. and Med., 85(1954):72-74.

13. Maloney, C. J., Calculation of median lethal dose when doses are subject to Poisson errors, Unpublished.

14. Haley, David C., Estimation of the dosage mortality relationship when the dose is subject to error. Technical Report No. 15, (1952), Applied Mathematics and Statistics Laboratory, Stanford University.

the human mortality rates DRA_h and DRB_h, and the animal mortality rates DRA_a and DRB_a.

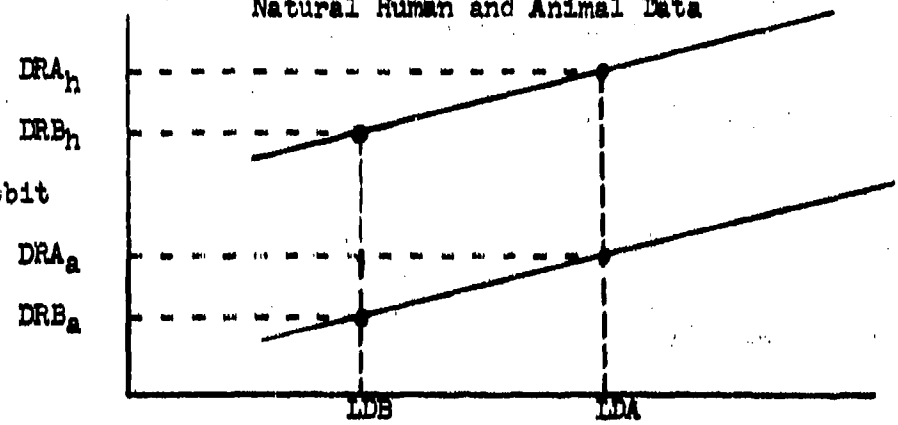
2) For the same animal species, conduct a laboratory bioassay experiment using animals trapped at the location under study, and calculate the dose-response animal mortality probit line.

3) Using the animal mortality probit line, compute the doses at levels A and B which correspond to DRA_a and DRB_a. Call these LDA and LDB.

4) Plot DRA_h vs. LDA and DRB_h vs. LDB on probit paper and connect the points with a straight line. This is an estimate of the human dose-response mortality curve.

Figure 7

Hypothetical Human Dose Response Determination Using Natural Human and Animal Data



VI. REMARKS. It is obvious that the preceding three "pure" methods do not exhaust the various possibilities, and that "mixed" procedures may be employed, or that several methods may be used and then combined to get a stronger overall estimate than that offered by the separate procedures.

On the other hand, these methods are only applicable provided the route of infection in nature is the route of interest.

The hypothesis of parallel morbidity and mortality probit lines can be tested on humans utilizing ideas from methods I, II, and III, if we can obtain in the field human morbidity and mortality data for several doses, and if we can either measure these doses directly or infer them from animal responses, as in methods II and III.

VII. INDEPENDENT ACTION MODEL. The probit transformation of the dose response curve outlined in Section II and discussed in detail in Finney, is

not the only one which has been proposed. Berkson[15] has suggested a model based on the use of the logistic curve rather than the integrated normal. Little practical difference exists between these two methods[16]. An alternative suggestion with enormous practical importance has been proposed apparently independently by Goldberg[17] et al. and by Peto[18]. The suggestion had previously been applied to transmission of plant virus[19], and recently in connection with the biological effects of radioactivity[20]. An early treatment was given by Yule[21]. A maximum likelihood procedure for fitting this curve has been given by Chernoff and Andrews[22]. Peto[23] has shown that, if this model fits, then every probit curve will have a slope whose numerical value is too irrespective of the agent, host, and route of administration. It is clear that if this model is correct then nothing is required for a complete determination of the dose response relation desired but the collection of cause specific death rates.

This consequence of the independent action model is so important that it is essential to determine whether or not the theory is substantiated. An estimate of the extent of experimentation required to provide tests of this hypothesis has been furnished Aerobiology Branch at their request[24].

15. Berkson, Joseph, Application of the logistic function to bioassay. J. Am. Stat. Assn., 39(1944):357-365.

16. Haley, David C., op. cit.

17. Goldberg, L. J., H. M. S. Watkins, M. S. Dohvatz, N. A. Schlamm, Studies on the experimental epidemiology of respiratory infections IV. Relationship between dose of microorganisms and subsequent infection or death of a host. J. Inf. Dis., 94(1954):9-21.

18. Peto, S., A dose response equation for the invasion of microorganisms. Biometrics 9(1953):320-335.

19. Watson, M. A., Factors affecting the amount of infection obtained by aphid transmission of the virus by Hy. III. Phil. Tran. Roy. Soc. 226, pp. 457-489.

20. Kimball, Allan, The fitting of multi-hit survival curves, Biometrics 9 (1953):201-211.

21. Yule, G. Udny, On the distribution of deaths with age when causes of death act cumulatively and similar frequency distributions. J. Roy. Stat. Soc. 73(1910):26-38.

22. Chernoff, Herman and Fred Andrews, A large sample bioassay design, Tech. Rpt. No. 17, Applied Mathematics and Statistics Laboratory, Stanford Univ.

23. Peto, S., op. cit. (pp. 329 ff)

24. Statistics Branch Job No. 1699, Dose response equation for microorganisms. Experimenters Dr. Persichetti and G. Broadwater. Statistician SP3 Richard Lamm, 1955.

SOME STATISTICAL ASPECTS OF FATIGUE TEST PLANNING

V. A. Didio
Watertown Arsenal

In studying metal fatigue one is usually most interested in the accumulated damage produced in a part or specimen that is subjected to repeated stresses. This is studied experimentally by subjecting specimens to repeated cycles of constant stress or constant deflection and observing the number of cycles at which failure occurs.

The stress applied to the specimen may be due to different types of loads, such as a compressive or tensile load, bending, torsion, or a combination of such loads; and, even though a constant stress is applied to a number of like specimens under as uniform a set of conditions as possible, there is observed considerable scatter in the number of cycles at failure, where failure can be defined in various ways. It could be considered as fracture, or the experiment could be stopped and failure said to have occurred in the specimen when some predetermined decrease in stiffness is observed.

Scatter is inherent in the experimental results due to the nonhomogeneity of the material on the microscopic and sub-microscopic scale and localized textural differences such as machining and heat-treating effects. The careful experimenter tries, insofar as he can, to eliminate the possible causes of these variations by standardizing techniques in preparing specimens. He makes the specimens from the same bar, or at least the same heat. He heat treats specimens under uniform conditions. He machines and polishes specimens such that residual stresses will not be introduced. In addition, variability due to the fatigue machine and its loading is reduced as much as possible. All of these parameters lead to varying life spans for individual specimens, as well as causing fracture at different positions along the specimens.

In spite of all these precautions, the life of one specimen differs from that of the next such that results of fatigue tests show a much wider scatter than the results of any other mechanical test. Even if the metal or alloy were free of all impurities or imperfections, a variability in its strength values would exist throughout its volume, because of its crystal structure. Variability cannot be eliminated and the scatter inherent in fatigue tests is accepted as a basis for the need of statistical analysis. The variation in number of cycles to failure of apparently similar specimens subjected to the same level of repeated stress obscures the results of many fatigue testing programs and makes it necessary to run a relatively large number of tests in order to obtain the desired information.

Theoretical explanations of the internal processes in the specimen which lead to failure are many and varied. They range from consideration of atomic dislocation movements to gross slip in individual crystals. Any attempt at a theoretical explanation of as complex a phenomenon as fatigue must necessarily appear as an over-simplification of the behavior of real materials.

The results of fatigue tests are usually presented in the form of S-N diagrams or curves; i.e., the stress S is plotted vs. the number of cycles N to failure. Usually these diagrams are determined by a rather arbitrary process of curve fitting through a relatively small number of points, which represent results of individual fatigue tests performed at several stress levels. They are obtained as "lines of best fit", in which case they are assumed to refer to the average fatigue performance of the specimens. Such presentation of the fatigue tests is necessarily inadequate, since it neglects a very significant aspect of all fatigue data, their scatter. Another deficiency of these S-N diagrams is that they are valid only within the range of stress amplitude under the repeated application of which a specimen actually fails, while our interest may be elsewhere.

Because of the significance of the scatter and its expected variation with the applied stress amplitude, results of fatigue tests can be effectively presented only by a relation between the stress, S, the number of cycles, N, and the probability, P, that any specimen subject to N repetitions of the stress amplitude, S, will actually break at or before N cycles (mortality function) or the probability that it will survive this number of cycles (survivorship function). This can be accomplished with statistical techniques, thus making a fuller use of the information present in fatigue results, while presenting them in a manner that is more meaningful and accurate.

In the design of structures and machine parts which will be subjected to loads and vibratory stresses, a reasonable safety from fatigue failures must be ensured, indicating a special concern with very small probabilities of failure or large probabilities of survival. These cannot be found directly by experimentation without testing a very large number of specimens. Therefore, results of fatigue tests are useful only if combinations of (SN) can be predicted by extrapolation, at which the probability of survival can be made as close to unity as desired with respect to the specified factor of safety. Such extrapolation beyond the range of the actual experiment, however, requires an adequate knowledge of the character of (SN) probability surface, and thus of the statistical distributionof N for constant values of S, as well as S for constant values of N, particularly in the vicinity of the endurance limit where the probability of survival approaches one.

We are thus left with finding a mathematical approximationof the fatigue phenomena as expressed through data collected by experimental studies. If we suppose the existence of an exact relationship between the life of a specimen and the stress to which it is subjected, and approximate it by some mathematical expression, it will be readily found that even if the approximation is not very close the number of tests necessary to reveal the difference between the exact and approximate relation will be surprisingly large, owing to the wide scatter present in the observed fatigue lives. An improvement of the approximation,either by changing the function or by increasing its number of parameters, will soon bring us to a position of being unable to decide experimentally whether or not there are any divergences. As a consequence, there may exist two or more relationships of different shapes that satisfactorily represent the data. Therefore, the only reasonable way to act seems to be to choose a function which most easily gives answers to posed questions and is still consistent with known fatigue properties of materials.

A distribution of fatigue life of specimens subject to a given stress amplitude that represents the actual distribution of test results rather closely is obtained by assuming that, in each large group of specimens tested at the same stress amplitude and subject to a number of load cycles, the specimen that actually fails at this number is necessarily the weakest specimen. Hence, the specimens that fail at various numbers, N, of load cycles may be considered as forming a group of the weakest specimens out of (large) samples of the population tested; to the analysis of the distribution of N in this group, the theory of extreme values might therefore be applied. The distribution of extreme values can thus be derived from any reasonable assumption concerning the distribution of the population from which the extremes are drawn. It must be noted that the use of the extreme value distribution has its strongest justification in that it is, as far as can empirically be established, a good approximation to actual test results.

As in most experimental investigations, the probability functions are actually determined from the test results. The direct determination of the frequency distribution would require a much larger number of experiments than can usually be performed. There is also available an extremal probability paper on which a graphical indication may be obtained concerning the possibility that a variable has an extreme value distribution. This would be shown by a straight line relationship between the variable and a reduced statistical variate, similar to the use of normal probability paper.

Under certain assumptions concerning the theoretical processes that produce fatigue, the fatigue life of the population at a particular stress level can be shown to be logarithmic normal, so that the distribution of $\log_{10} N$ in the population of specimens can be expected to be normal. This normality of $\log N$ was first noticed in the results of experimental tests. The specimens that actually break at given values of $\log_{10} N$ are thus the weakest specimens in samples of the normally distributed population of fatigue lives.

In dealing with the exact distribution of extremes, many difficulties are encountered in numerically evaluating it, even when the initial distribution is known. To overcome this obstacle, asymptotic distributions valid for large samples were derived[1]. These asymptotic distributions vary depending on the initial distribution from which the extremes were taken and whether or not the variate being considered is limited or unlimited in the direction of the extreme being considered. When the initial distribution is of the exponential type, as for example the Normal Distribution, we have the first equation, which is the asymptotic probability of the smallest value x. y is a reduced variate analogous to the standarized variate used in normal distributions. α is a measure of dispersion, and u is the mode of the distribution of x.

[1] Gumbel, E. J., "Statistical Theory of Extreme Values and Some Practical Applications"; N.B.S. Applied Math. Series #33.

(1) $$\Phi(x) = \exp\left[-e^{y}\right]$$

where

$$y = \alpha(x - \mu)$$

(2) $$P(x) = \exp\left\{-\left(\frac{x - w}{\mu - w}\right)^{k}\right\}$$

$$x \geq w;\ \mu > w.$$

The function represented by the second equation is for the extremes of smallest values, where k is a measure of dispersion and variate x now has a lower limit. This function can also be derived from the first equation by a logarithmic transformation of the variate and is known as the third asymptotic probability function.

Before analyzing our survivorship function, we must make assumptions concerning the existence of an upper or lower limit of the function, which must be based on experimental facts. This will determine our choice of analysis, for while we usually are more interested in the endurance limit of a specimen there is also the problem of whether or not there is a minimum life for this sample, i.e., a certain N at a stress level, S, such that failure will not occur below this number of cycles. In fatigue tests, the stress is kept constant and the number of cycles to failure N noted for each group of specimens, implying that such an N exists and is greater than zero, although this may not be true for soft metals. Thus, since our variate N is limited, we use what is referred to as the third asymptotic probability function.

Our design of the experiment will also depend on what particular aspect of fatigue we are interested in--endurance limit, minimum life, or median fatigue life. This will determine the placement of the various numbers of stress levels that we will use. The stress levels should be sufficiently far apart to make the test results significantly different, but near enough to allow us to construct a survivorship function.

In particular, the object of most fatigue programs is the determination of the endurance limit of a specimen or part. The true endurance limit is the greatest stress for which the probability of surviving an infinite number of cycles equals one. The estimate of the true endurance limit cannot be checked, since we cannot let the testing machine run for an indefinite number of cycles. For this reason, it has been customary in testing steels to replace infinity by 10^7 cycles andto define endurance limit as the largest stress for which the probability of surviving 10^7 cycles at this stress is one. At this stress level, failure becomes independent of N--that is, the (SN) curves become parallel to the N axis.

Estimation of the endurance limit based on a specific interpretation of the existing data by using probabilities of survival and statistical theory. If an analytic expression for S as a function of N could be derived from physical considerations, its extrapolation for $N = \infty$ would lead to knowledge of the endurance limit. Since no such expression is known, the endurance limit stress has to be estimated by extrapolation from the probability of survival valid for large values of N. For this purpose, we need a specific distribution theory.

Theoretical considerations such as those which led to the extreme value distribution can lead us to approximations of the observed physical phenomena. These approximations must be tested against experimental results, experimental results which are sufficient and accurate to enable us to arrive at some conclusions or to give us indications on how close our approximation stands to reality. Our need now is for verification or alteration of our premises by experimentation.

The tool for our estimation of the endurance limit is the probability of permanent survival, which is a function of stress. This probability will be estimated from the number of specimens that failed and the number that survived at different stress levels. These will be analyzed with the help of the asymptotic theory of smallest values of a non-negative variate and, in turn, will lead to an estimation of the endurance limit. Available probability tables [2] for the analysis of extreme-value data aid us in determining our parameters and in estimating the endurance limit.

The probability of permanent survival is usually determined from experiments performed in the following manner:

A number of specimens is subjected to a constant maximum stress during an increasing number of cycles, N, up to failure. The number of cycles at failure, N_f, is recorded, or if no failure occurs the experiment is stopped at a high number, say $N = 10^7$ or 10^8. Specimens are usually tested first at a stress level such that either all specimens fail or a small proportion survives. This experiment is then repeated for a number of different stresses. From these results, we can determine the probability of survival as a function of the variate N for each value of S tested, noting that for constant N the probability of survival increases as the stress decreases.

These probabilities could also be determined by subjecting a number of specimens to a fixed stress, S, and stopping the experiment at a predetermined number of cycles, noting the proportion of survivors at each stress. This would be repeated for the same number of cycles at lower and higher stresses, such that the range of variation of the stress reached from the low stress where all specimens survive up to the high stress where all specimens fail for the same number of cycles. These results would enable us to determine the probability of survival as a function of S for constant number of cycles, N, where the stress, S, now takes on the role of a statistical variate, although it is constant within each experiment.

[2] "Probability Tables for the Analysis of Extreme Value Data"; N.B.S. Applied Math. Series #22.

Thus, for a constant value of the probability of survival, there corresponds a series consisting of different numbers, N, as a function of S, or S as a function of N, such that, if N is plotted on the abscissa and S on the ordinate, (SN) curves are obtained where S decreases for increasing values of N for any constant probability.

These three representations of fatigue data are linked; each one must be compatible with the other two, making it unacceptable to use an empirical relation for one of these functions if it contradicts the theoretical properties of the other two functions. Also, any conclusion drawn from an alleged discontinuity of an (SN) curve must be wrong, since there is no reason to doubt the continuity of the survivorship functions.

In Figure 1 we have a schematic (SN) diagram where each curve corresponds to a fixed probability of survival. The top curve corresponds to a small probability of survival and, for all combinations of S and N above this curve, failure is practically certain.

The middle curve is for a probability of survival 1/e = 0/36788. The S and N values for this curve are called the characteristic stresses and number of cycles to failure, respectively. These values arise when y = 0 in Equation (1).

The lowest curve in Figure 1 consists of S and N values, before which no failure occurs. From this curve we can find our endurances at any number of cycles. For values of S and N below this curve, survival is certain in a probability sense.

Notice that the (SN) curves become parallel to the N axis as N approaches 10^7. The stresses at this number of cycles are used in estimating our true endurance limit stress, which we have defined as the largest stress for which the probability of surviving 10^7 cycles is one.

Our discussion has centered about the tail end of the distribution under study, since establishment of an endurance limit is important to continued studies of variables that appear in fatigue testing; but such information is no better than the knowledge of its accuracy. This is not only a statistical problem but also one of lack of sufficient fatigue data adequate for statistical interpretation, which shortage should be alleviated since its existence is now so evident.

There is still much research needed on the characteristics and behavior of various extreme value distributions. We especially must increase our knowledge of their behavior for small sample sizes. Also, the optimum number if specimens that should be tested at any stress level is still a matter to be decided, due to our lack of knowledge of the distribution of our estimation of the parameters.

Knowledge of the parameters, their distribution, the effect of sampling errors, confidence limits, etc. are essential before we can start on another aspect of fatigue testing, which should be the ultimate aim of the study of fatigue, i.e., the determination of a theory for predicting the behavior of

materials under repeated stress. There are various theories seeking to explain fatigue from the viewpoint of engineering principles. These theories develop under controlled experimentation and achieve what validity they have by being statistically significant and physically consistent.

Successful study of such variables as position of failure, effect of size and shape, the frequency of load cycles, temperature--all are dependent to various degrees on the determination of the endurance limit.

Here, statistics has a two-fold job.

It must aid the design engineer in the design of parts or machines by giving him a criterion concerning fatigue life or endurance limit on which to base his analysis.

Then, it must develop as a tool which will enable the engineer and metallurgist to better understand the phenomena which is now referred to by the general term, brittle behavior, and allow him to evaluate the effects of introduced variations on metals.

BIBLIOGRAPHY

(1) Gumbel, E. J., "Statistical Theory of Extreme Values and Some Practical Applications". N.B.S. Applied Math. Series #33.

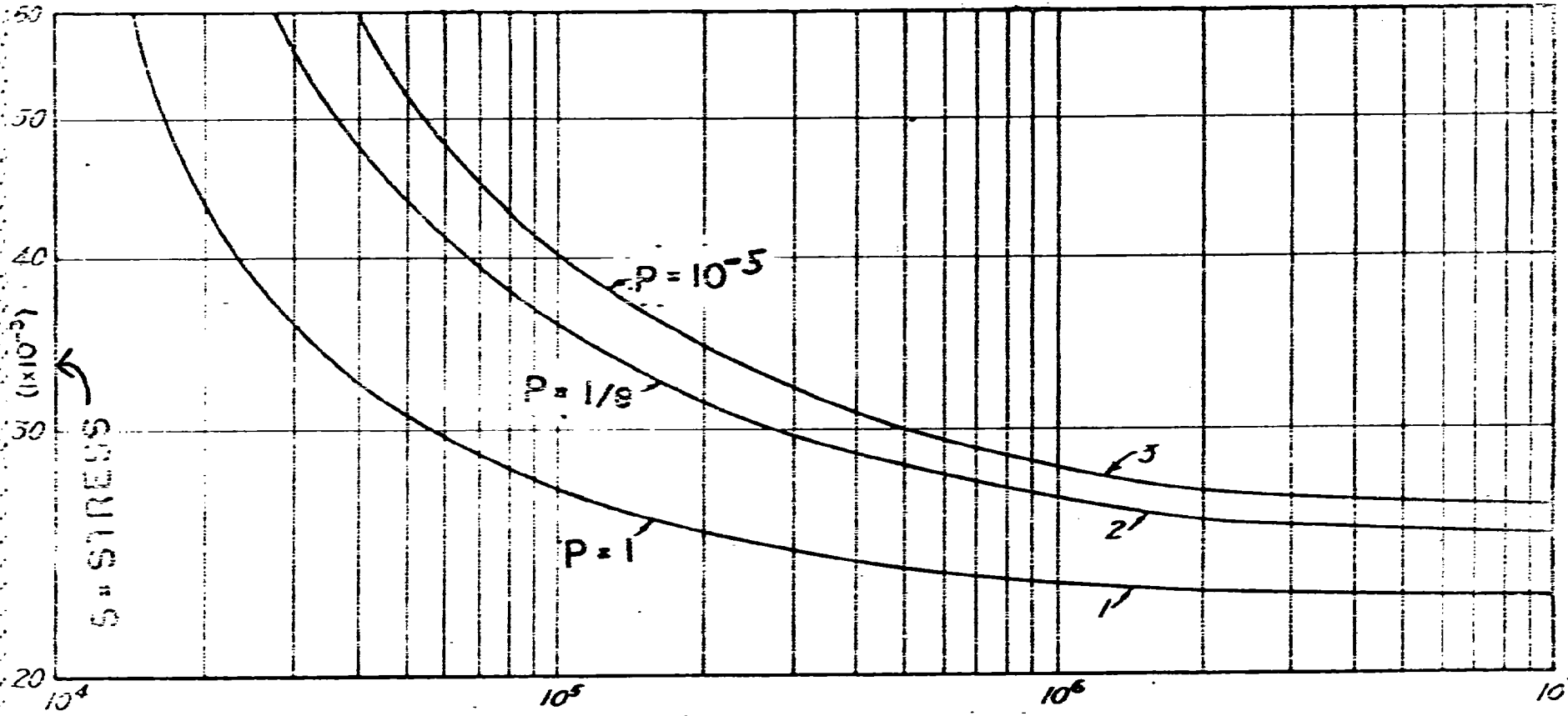
(2) Probability Tables for the Analysis of Extreme-Value Data. N.B.S. Applied Math. Series #22.

(3) Freudenthal, A. M. and Gumbel, E. J. Proceedings of the Royal Society, A, Vol. 216, London, 1953.

(4) Gumbel, E. J., "Statistical Estimation of the Endurance Limit"; Technical Report T-6A, Office of Ordnance Research.

(5) Lieblein, J., "A New Method of Analyzing Extreme-Value Data"; NACA Technical Note 3053.

SCHEMATIC (SN) DIAGRAM



N = NUMBER OF CYCLES AT FAILURE

FIGURE 1

THE USE OF A SPECIAL SYSTEMATIC DESIGN FOR SURVEILLANCE TESTING

Robert M. Eissner
Ballistic Research Laboratories

In testing field artillery ammunition in order to evaluate its ballistic quality, i.e., its exterior and interior ballistic characteristics, range and velocity, separate loading ammunition presents us with a different problem from that involved in testing fixed or semi-fixed ammunition. In fixed or semi-fixed artillery ammunition, or any artillery ammunition as that matter, we have what is called a complete round. This complete round is composed of a fuze, a projectile, a propellant, and a primer, all of which are packaged, stored and issued as one unit and can be loaded into a weapon in one operation. A group of these units with certain restrictions imposed upon it, e.g., being manufactured under similar conditions within certain time periods, using only one propellant lot, using not more than two primer lots and fuze lots, and using empty projectile lots from only one manufacturer, comprise a complete round lot. When a sample from this complete round lot is fired in the field, it can be said that the measured range and velocity are characteristics of that one, and I repeat one, complete round lot. Thus each complete round lot as such in storage has a range and velocity. However, with separate loading ammunition such is not the case. As might be suspected from the name, each of the components, namely the shell and the propellant, are packaged, stored and issued separately and are also loaded into a weapon separately. Thus, since any propellant lot might be fired with any number of projectile lots in the field and vice versa, the concept of a measured range and velocity for each complete round lot of separate loading ammunition in storage does not exist. The propellant as a separate item of issue has its characteristic, velocity, and the shell as a separate item of issue has its characteristic, range.

Now in surveillance testing it is desired that the quality of each lot in storage, whether it be a lot of fixed ammunition, semi-fixed ammunition, separate loading projectile or propellant, be evaluated. To do this, periodically lots of a given type of ammunition are sampled and fired in some manner in order that those characteristics range, velocity, functioning, etc., which are needed to ascertain the quality of a lot may be obtained. Upon obtaining these characteristics, say mean range, standard deviation in range, mean velocity, standard deviation in velocity, number of duds, number of low order functionings, etc., a lot may be assigned one of four grades by using a set of previously established Lot Quality Standards. Thus, in this manner the quality or grade of the individual lots in storage may be evaluated. However, in addition to this it is also desired that over-all estimates of the round-to-round and lot-to-lot dispersions for a particular ammunition type be obtained. Such information is of great benefit to the using field forces, those people involved in preparing firing tables, and those people involved in weapons systems analyses. With this brief description regarding surveillance testing of artillery ammunition, the problem involved in testing separate loading ammunition may be clearly seen, that is, how can we fire an economically feasible test and still get the desired results mentioned previously?

In answering this question it must first be realized that it is most difficult, in fact almost impossible, to control all the extraneous factors that may affect a ballistic test. In no way is a ballistic test like a laboratory experiment where most of the factors can be rigidly controlled. Weather conditions, tube conditions, etc., once a test has started just cannot be controlled. Consequently, in a surveillance test, we must be sure that we're getting the unbiased estimates of the parameters needed to grade a lot, i.e., that we're getting estimates that actually reflect differences in lots and not differences due to methods of test or other extraneous factors. We want to be sure that we will not penalize or downgrade any lot for any other reason than inferior performance. For these reasons then it is necessary that we make use of designed experiments and/or reference lots or standard lots as they are often called. In this way we hope to eliminate or minimize any extraneous factors and to estimate the parameters for each lot with equal precision. Having all this information, there are two general methods of test that can be employed in order to get the desired results--those in which test propelling charge lots and test shell lots are fired in the same program and those in which test propelling charge lots are fired with reference shell lot (the reference shells all being loaded to the prescribed standard weight) in one program and test shell lots are fired with the reference propellant lot in another. One word here on what is meant by a reference shell lot or a reference propellant lot. A reference lot is that lot which has been standardized and fires a known or firing table value when fired under standard conditions, i.e., standard meteorological conditions, new gun tube, standard propellant temperature, etc. Generally extensive firings using a number of different tubes on each of several days have been conducted on these reference lots in order that the greatest possible amount of information about the lot is available. Now getting back to the methods of test, the first method, the one in which the test propelling charge lots and the test shell lots are fired in the same program, is greatly more economical. In fact it involves only about half as much firing as does the second method. In addition it also more nearly approaches actual field firing conditions, where a mixture of propelling charge lots and shell lots may be fired during the same mission although they are not supposed to be fired in that manner. The second method, however, is a less complicated procedure and gives estimates of mean range and/or velocity and standard deviation of range and/or velocity better suited for surveillance purposes, i.e., grading of the individual lots. It is also the procedure generally followed in the acceptance tests of the ammunition.

Now that we have given these general descriptions of the two methods, let us discuss them in more detail. For programs of the first type, various combinations of the different shell lots and the different charge lots are made into complete rounds as defined previously. Included among the different shell lots is the reference shell lot and included among the different propellant or charge lots is the reference propellant lot. These reference lots enable us to tie in the results from this test with those from previous or future test. They serve as a control lot and theoretically take out any day-to-day or occasion-to-occasion effects. Getting back to the design, two complete rounds from each of the possible combina-

tions of charge lots and shell lots--by this I mean each propellant lot is combined with every one of the shell lots and vice versa--are fired in pairs as a two factor experiment. Diagrammatically the design for, say, four test lots of shell and four test lots of propellant looks something like this:

Propellant Lot \ Shell Lot	Ref. Shell	Lot 1	Lot 2	Lot 3	Lot 4
Reference Propellant	1a, 1b	6a, 6b	11a, 11b	16a, 16b	21a, 21b
Lot A	22a, 22b	2a, 2b	7a, 7b	12a, 12b	17a, 17b
Lot B	18a, 18b	23a, 23b	3a, 3b	8a, 8b	13a, 13b
Lot C	14a, 14b	19a, 19b	24a, 24b	4a, 4b	9a, 9b
Lot D	10a, 10b	15a, 15b	20a, 20b	25a, 25b	5a, 5b

The number shown in the cells refer to the sample round number. For example, sample rounds 1a and 1b consist of the refesrence shell and reference propellant, sample rounds 2a and 2b consist of shell from lot 1 and propellant from lot A, etc. Regarding the order of fire, the first group of ten rounds (Nos. 1a thru 5b) are fired first followed by the second group of ten rounds (Nos. 6a thru 10b), etc. until all five groups of ten rounds are fired. Within each group of ten rounds, however, the sets of two samples are fired in a random order. For example, the first group of ten rounds could be fired as follows: 3a, 3b, 5a, 5b, 2a, 2b, 1a, 1b, 4a, 4b.

At first it was intended to fire the program as a Latin Square. As you can see, shell lots, propellant lots, and order of fire would be the three factors. However, the order of fire for any groups of pairs was randomized thus destroying one of the underlying conditions of the Latin Square design--that each treatment occurs once and only once in each row and each column. This was done in order to preclude any possibility of a memory effect that may come about from an ordered design. To digress once again by memory effect is meant the effect on lot B due to the fact that it always follows lot A in the firing sequence. These memory effects, which usually invalidate the data for a program, are constant hazards in any ballistic test since they may be caused by any number of seemingly unimportant factors, for example, small changes in the chemical composition or web size of the propellant. Two classic examples of such memory effects occurred in the 90mm gun. One case occured during World War II and was caused by the addition to the propellant of a small amount of potassium sulfate which had been added to suppress flash. The effect of this small change was that when a sulfated propellant and a non-sulfated propellant were fired alternately in a relatively new tube the non-sulfated rounds were depressed from the normal by about 20 f/s in velocity whereas the sulfated rounds fired correspondingly higher. Since propellants are accessed in this manner, i.e., alternately firing the test propellant and the standard propellant, the assessment or the charge weight that will enable the propellant to fire the required or service velocity of many 90mm non-sulfated propellant lots was in error by about 40f/s due to the

fact they were assessed against a sulfated reference propellant. The second case occurred during the Korean War and was very similar in nature. It involved a 10% change in the web size of the propellant. The web size of the test propellants was increased by 10% whereas the web size of the standard propellant was not changed. This too resulted in approximately a 40f/s error in velocity for the test propellants. In case you're interested both situations were remedied quickly by standardizing a new reference propellant which had the same physical properties as the test propellants being produced.

Now getting back to our discussion, for programs of the second type, the different propellant lots are assembled into complete rounds with the reference shell lot when propellant lots are being tested and the different shell lots are assembled into complete rounds with the reference propellant lot when shell lots are being tested. These complete round lots are then fired in a series of five round groups in a manner determined by the number of lots being tested. For example, if three test lots are being tested the firing sequence would be reference lot, test lot 1, test lot 2, test lot 3, reference lot; if four tests lots are being tested the firing sequence would be the same as that above except that four groups of test lots would be fired between the reference groups; if six test lots are being tested the firing sequence would be reference lot, test lot 1, test lot 2, test lot 3, reference lot, test lot 4, test lot 5, test lot 6, reference lot; etc. In each of these cases the sequences would be fired a second time in order that ten rounds from each test lot would be fired.

In firing each of these designs certain other control mechanisms are used in order to minimize any extraneous effects that would bias the results. These mechanisms include the use of only one gun tube throughout the program, storing the ammunition at a constant temperature of 70°F for approximately 24 hours prior to firing, firing conditioning rounds of the same type and composition as the test rounds before any of the test rounds are fired in order to get the gun tube in the proper frame of mind so to speak, using the same lot of fuzes and the same lot of primers throughout the program, and firing any one phase of the program on one day without cessation or any undue delay.

With this description of the practices and procedures involved in the ballistic testing of ammunition you have become acquainted with two methods of testing separate loading ammunition--that method which we shall call Method 1 where test charge lots and test shell lots are fired in various combinations in the same program as a two factor experiment and that program which we shall call Method 2 where the test shell lots are assembled with the reference propellant lot and fired in one program and the test propellant lots are assembled with the reference shell lot and fired in another. The first method better simulates field firing conditions and is more economical whereas the second method is more easily accomplished and gives results better suited for surveillance purposes.

In order that we may make a comparison of the two methods of test a program has been fired involving four test lots of M1A1 155mm Howitzer

propelling charges and four test lots of HE M107 155mm Howitzer shell. Ten rounds from each of the test lots were fired in each of the three zones, III, V, and VII. In firing Method 2, however, only that phase involving the firing of the test propelling charge lots with the reference shell lot was conducted. For this reason then only the characteristic muzzle velocity is considered in making the comparison. Comparing the results of the two methods after analyzing the data from each we have the following:

CHARGE III

	Method 1	Method 2
Avg. Vel	873.0f/s	869.0f/s
Rd-to-Rd Std. Dev.	7.98 f/s	5.20 f/s
Lot-to-Lot Std. Dev.	6.71 f/s	5.03 f/s

CHARGE V

Avg. Vel.	1223.2 f/s	1218.5 f/s
Rd-to-Rd Std. Dev.	4.47 f/s	4.00 f/s
Lot-to-Lot Std. Dev.	2.87 f/s	1.65 f/s

CHARGE VII

Avg. Vel.	1848.2 f/s	1844.8 f/s
Rd-to-Rd Std. Dev.	4.10 f/s	3.02 f/s
Lot-to-Lot Std. Dev.	1.11 f/s	2.11 f/s

In each of the charges it is observed that the average velocity of the lots obtained from the first method is larger than the average velocity of the lots from the second method. In fact in each case the average velocity obtained using method one is significantly greater. It is likewise observed that the round-to-round standard deviation in velocity obtained from the first method is greater than that obtained from the second method. In this case, however, only in Charges III and VII is the round-to-round standard deviation obtained from the first method significantly greater. In no case are the lot-to-lot standard deviations significantly different.

Having observed these results the question comes to mind why are the results from the two methods different? Just why should method one give larger round-to-round dispersions than those of the more commonly used second method? In an attempt to answer this question we will further analyze the first method since by the nature of its design, as opposed to the simplicity of the second method, it more readily lends itself to extensive analysis. Analyzing it first as a two-way classification with two observations per cell it was observed that in all three charges there was a highly significant shell and propellant interaction effect. This was rather surprising since the test had been designed under the supposition that any such effect would be negligible. To investigate the possible causes of this interaction effect and also possibly throw some light on the differences in the results for the two methods, we made several corrections to the data. These corrections were made to account for known differences between the two methods. The first correction made was that for differences in shell weights. It's remembered that in the second method reference shell all loaded to the prescribed standard weight are used whereas in the first method test shells

loaded to various weights are used. Thus correcting each velocity for the variation of the shell weight from the standard weight would take out any effect due to shell weight. Making this correction we found, as expected, had no significant effect, in fact hardly any effect at all, on the results of the two-way classification. The second correction was that for velocity trend. As more rounds are fired from a tube the velocity level of the tube usually becomes lower. This is generally more true of high velocity weapons and is not considered of too great importance when firing the smaller caliber howitzers, especially when firing only a fifty round group. However, since we are interested in investigating all the possibilities, we estimated the velocity trend using the analysis of covariance and then removed any trend found from the data. Doing this reduced the interaction effect in each case and in some cases even made it insignificant. Based on this result then the velocity trend evidently did cause some of the interaction. However, neither it nor the shell weight correction had any effect on the round-to-round standard deviation and very little effect on the average velocity.

Thus in view of these results no light can be shed as to the reasons for the larger dispersions and higher velocities of the first method other than that of the difference in the experimental errors in the two test procedures. Therefore, unless some physical means of evaluating the magnitude of this difference is obtained, the only way the first method can be used in order to assign grades to the individual lots without unnecessarily penalizing them is to have the Lot Quality Standards and Criteria take into account such increases and be based upon experimental data from tests of the first type. In this way then the more economical first method could be used and individual lot grades could still be assigned.

To summarize, having given you a brief description into the difference between separate loading and fixed and semi-fixed ammunition and also having given you the main purposes of surveillance testing, that of grading individual lots and providing over-all estimates of dispersion for different types of ammunition, you were made aware of the problem involved in surveillance testing separate loading ammunition--how to economically and realistically test separate loading ammunition and still get results that may be used to achieve the purposes of surveillance testing. To accomplish this, because of the many extraneous factors that may affect ballistic tests, the use of designed programs and reference lots had to be used. Two such kinds of programs were given: program or method one involved firing test propelling charge lots and test shell lots in the same design, whereas program or method two involved firing the test propelling charges lots with the reference shell in one phase and the test shell lots with the reference propellant in the other. Programs of the first type were more economical and more nearly characterized the manner in which separate loading ammunition was fired in the field; programs of the second type gave results which were better suited for grading individual lots. The results from a program comparing the two methods were given. These results showed that programs of the first type gave in most cases significantly larger round-to-round standard deviations and significantly greater average velocities. No explanation for these increases was found although velocity trend appeared to play a significant role with respect to the interaction. Therefore,

based on the findings of the special program, it was concluded that the only way in which the more economical and realistic first method could be used in order to assign grades to the individual lots without unnecessarily penalizing them was to have the Lot Quality Standards and Criteria take into account such increases in experimental error and be based upon experimental data from tests of that type.

A STATISTICAL DESIGN FOR A SURVEILLANCE TEST

Boyd Harshbarger*
Redstone Arsenal and Virginia Polytechnic Institute

An example may serve to show how the problem of surveillance can be attacked through statistical design. We will discuss a portion of a well-designed experiment carried out by the Rocket Development Group at the Redstone Arsenal. The variable concerning us in this talk is the time to spontaneous ignition in the sample tested. This was one of several variables measured in the study. The other variables were strand burning rate and X-ray diffractometric analysis of oxidizers on the surface of the sample. A study was made on the sizes of the variances and means before and after running each test. This study served to detect a shift in the means as well as to measure variability due to the techniques, equipment, and personnel.

The observations follow the usual linear model,

$$y_{ijkh} = \mu + \alpha_i + \gamma_j + \delta_k + (\alpha\gamma)_{ij} + (\alpha\delta)_{ik} + (\gamma\delta)_{jk} + (\alpha\gamma\delta)_{ijk} + \varepsilon_{ijkh}$$

where μ is the overall mean, α_i is the added effect of the i^{th} sample, γ_j is the added effect of the j^{th} week, δ_k is the added effect of the k^{th} environment, $(\alpha\gamma)_{ij}$ is the added effect of the interaction of the i^{th} sample, and the j^{th} week, $\ldots$, and the ε_{ijkh} are random errors, independently and normally distributed with zero means and common variance σ^2. The important things to observe here are that we are dealing with fixed or named effects, that our model is a linear one and that the model includes a factorial. In a factorial experiment, the effects of a number of different factors as well as their independence are investigated simultaneously. In reality the environments are further separated into helium and oxygen and each at two different temperatures. All this modification does to the model is to add several terms.

It is easily shown that the least square solution of the linear model gives estimates of the various effects and also provides the basis for an analysis of variance. This analysis of variance provides a test of signifigance in which one compares the random error with the treatment and interaction effects.

The chemists see the objectives of the experiment as:

(a) To compare the behavior of the basic samples designated as D and U over a period of time.

(b) To establish the week-to-week trend, if it exists.

(c) To compare the effects of two different tests of environments, helium and oxygen.

(d) To investigate the effects of temperature.

* The author acknowledges the help of Lt. E. L. Bombara and the supplying of the data by Mr. R. L. Rudolph, both of Redstone Arsenal.

(e) To study the interaction or independence of the main effects.

The data that were gathered to answer these questions are given in Table I.

The statistician attempts to show a mathematical model and analyses which will enable the chemist to answer his questions on a probability basis. In general, this involves the setting up of a number of so-called null hypotheses, which may or may not be rejected.

Table I gives the time in seconds to spontaneous ignition for the samples tested.

TABLE I

Time to Spontaneous Ignition for the Tested Samples

	U				D			
	ENVIRONMENTS				ENVIRONMENTS			
	Helium		Oxygen		Helium		Oxygen	
	a 70°	b 120°	c 70°	d 120°	a 70°	b 120°	c 70°	d 120°
	89.1	89.4	89.7	75.2				
AFTER	86.6	94.5	85.9	73.4				
ONE	89.9	93.7	86.8	74.5	88.8	92.8	86.2	84.7
WEEK	85.5	90.0	91.0	77.5	94.1	92.2	92.8	83.6
	87.7	87.8	84.4	72.8	90.0	92.3	90.4	79.8
$\bar{X}$	87.8	91.1	87.6	74.7	91.0	92.4	89.8	82.7
	86.7	92.1	85.3	76.6				
AFTER	89.4	88.9	84.3	73.6				
TWO	84.5	90.4	80.6	74.4	89.5	96.4	83.3	81.6
WEEKS	87.7	89.0	78.3	75.4	89.6	94.1	85.7	81.1
	87.1	90.5	79.9	72.5	93.4	98.4	87.4	79.4
$\bar{X}$	87.1	90.2	81.7	74.5	90.8	96.3	85.5	80.7
	90.2	86.9	83.4	76.2				
AFTER	88.7	84.3	82.3	71.8				
THREE	86.7	92.9	82.1	67.7	91.8	97.3	90.0	84.8
WEEKS	87.7	87.8	80.7	70.9	94.3	96.5	86.1	79.0
	88.1	85.1	83.0	77.4	86.8	96.9	86.5	82.6
$\bar{X}$	88.3	87.4	82.3	72.8	91.0	96.9	87.5	82.1
	90.2	81.6	79.7	67.3				
AFTER	90.5	89.1	77.2	65.5				
FOUR	88.1	87.4	78.9	66.1	95.5	91.1	86.5	71.6
WEEKS	87.3	86.9	82.9	62.9	94.2	91.1	81.9	73.1
	87.2	86.1	74.6	67.3	91.4	93.6	84.9	69.4
$\bar{X}$	88.7	86.2	78.7	65.8	93.7	91.9	84.4	71.4

The usual calculations on the data from Table I are now made to give the analysis of variance. Under the column, "source of variation," are shown the several types of variation and opposite these names, under the column headed "mean square," are given comparable estimates of these variations. The quantity opposite "error" under the mean square is an estimate of random variation. Comparison between the "error" and the other mean squares is used to produce a test of significance. Table II gives the analysis of variance.

TABLE II

Analysis of Variance of Time to Spontaneous Ignition

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
U vs D	1	817.20	817.20*
Weeks:	3	388.57	129.52*
Linear	1	325.76	325.76*
Quadratic	1	27.57	27.57*
Cubic	1	35.25	35.25*
Environments:	3	5116.17	1705.39
Temp's	1	504.83	504.83*
He vs O_2	1	3644.45	3644.45*
Temp's x He vs O_2	1	966.89	966.89*
U vs D x Weeks	3	34.53	11.51
U vs D x Environments	3	58.11	19.37
Weeks x Environments	9	444.96	49.44*
Weeks x Temps	3	172.30	57.43
Weeks x He vs O_2	3	256.80	85.60
Weeks x He vs O_2 x Temp	3	15.86	5.29
U vs D x Weeks x Envrs	9	64.94	7.22
Error	96	534.92	5.57
TOTAL	127	7459.40	

Means are presented in Tables III, IV, and V. Table I[I] gives some indication as to the significant trends of these tables and also indicates which trends can be dismissed as purely random variation.

Time in weeks seems to affect both samples, U and D, in the same manner. The environment, however, shows that they vary from week to week in a different manner for the separate conditions a, b, c, and d. Temperature affects the time to spontaneous ignition differently in helium than in air.

The week-to-week variations show a linear trend but not sufficiently that the remaining variation is non-significant. The two samples, U and D, gave different times to spontaneous conbustion. By looking at the analysis of variance table, one can see other variations that are significant.

TABLE III

Weeks

		I	II	III	IV	Avg
U vs D	U	85.3	83.4	82.7	79.8	82.8
	D	89.0	88.3	89.4	85.4	88.0
	Avg	86.7	85.2	85.2	81.9	

TABLE IV

Environments

		a	b	c	d	Avg
Weeks	I	89.0	91.6	88.4	77.7	86.7
	II	88.5	92.5	83.1	76.8	85.2
	III	89.3	91.0	84.3	76.3	85.2
	IV	90.6	88.4	80.8	67.9	81.9
	Avg	89.3	90.8	84.1	74.7	

TABLE V

Environments

		a	b	c	d	Avg
U vs D	U	87.9	88.7	82.6	72.0	82.8
	D	91.6	94.4	86.8	79.2	88.0
	Avg	89.3	90.8	84.1	74.4	

By extending the analysis of Table II, some revealing facts can be shown as indicated in Table VI.

TABLE VI

Source	d.f.	SS	MS	F
U vs D	1	817.19	817.19	146.71**
Environments	3	5116.17	1705.39	306.17**
Temp with He	1	37.21	37.21	6.68*
Temp with Oxy	1	1434.51	1434.51	257.54**
Gases (Helium vs Oxygen)	1	3644.45	3644.45	654.30**
Weeks within a	3	18.68	6.23	1.12
Linear	1	12.38	12.38	2.22
Quadratic	1	6.04	6.04	1.08
Residual	1	.26	.26	
Weeks within b	3	75.08	25.03	4.49**
Linear	1	50.06	50.06	8.99**
Quadratic	1	24.32	24.32	4.37*
Residual	1	.69	.69	

TABLE VI (con'd)

Source	d.f.	SS	MS	F
Weeks within c	3	241.87	80.62	14.47**
Linear	1	185.98	185.98	33.39**
Quadratic	1	6.94	6.94	1.24
Residual	1	48.95	48.95	8.79**
Weeks within d	3	497.91	165.97	29.80**
Linear	1	357.30	357.30	64.15**
Quadratic	1	113.63	113.63	20.40**
Residual	1	26.98	26.98	4.84*
U vs D x Weeks	3	34.53	11.51	2.07
U vs D x Envr	3	58.11	19.37	3.48**
U vs D x Week x Envr	9	64.94	7.22	
Error	96	534.92	5.57	
Total	127	7459.40		

In Table VI, the variation is separated so as to show separately the variation of weeks in the four different environments. Weeks within environment (a) is not significant, but when heat is applied to produce environment (b), a variation between weeks is noted. Weeks within environment (c), which is at ambient temperature and in oxygen, is greater than the variation between weeks within environment (b) but is still less than the variation noted for between weeks within environment (d) which is at the higher temperature. The pattern for this analysis of variance shown in Table VI is useful in many factorial experiments.

The analysis of variance was run on the logarithms of the estimated variances (s^2) calculated from the within sample variations for both sample U and sample D. There was no significance noted in either analysis variance of variances.

There may be some objection to considering the mean square with ninety-six degrees of freedom as an experimental error in as much as it has many characteristics of a sampling error. A more realistic experimental error

may be obtained by using the first and second interaction terms. This error would involve the interaction of weeks, environments, and the second order interaction of weeks and environments with the differences between the samples. It appears reasonable to assume that the interaction of weeks and environments with the differences between samples will be a random variable and thus given an estimate of true error. For Table VI the error term would be 10.50 with nine degrees of freedom. A chemist is primarily interested in the types of curves and the estimate of residuals from these curves. It can be seen that in the environment with helium, linear and quadratic trends account for most of the variation. In oxygen there is a different picture, as no specific trend appears and the significant variation between weeks is accounted for by the results for the last week.

MONTE CARLO AND OPERATIONAL GAMING IN ORDNANCE RESEARCH

L. M. Court
Diamond Ordnance Fuze Laboratories

It has been said that the proper role of a meeting chairman is to serve the needs of his audience; he should not obtrude on the speakers or the discussion but confine himself to listening. Briefly, he is a sort of program traffic director. If this is the case, then in submitting this "post mortem" comment on what went on at one of the sessions, the writer is sinning against the code of good conduct for chairmen. His only excuse is the importance of the topics to be touched on: Monte Carlo and Operational Gaming. This and the fact that operational gaming is the heart and substance of the first of the three papers presented under his chairmanship, and the fact that Monte Carlo is the technique that resolves the central problem of another paper.

Both Monte Carlo and operational gaming have burgeoned in the era since Von Neuman and Morgenstern wrote their classic on the theory of games and the modern high speed electronic computer became a practical operating device; indeed, Von Neuman himself, in company with another mathematician, Ulam, is responsible for the Monte Carlo idea in its modern version, as it is currently being exploited by physicists and operations research analysts, although the ancestry of the idea can be traced back at the very least to the time of Buffon and his celebrated needle problem. Allowing for the brief decade or so that Monte Carlo has been pursued, a not inconsiderable literature has grown up about it, although the bulk of the published material busies itself with actual examples rather than broad theory. Would-be enthusiasts, who recognize the power of the method but are otherwise uninitiated, justly complain that a satisfactory introduction is hard to come by. The truth is that Monte Carlo is in its sheerest infancy, and many problems remain to be resolved; e.g., what is the full gamut of mathematical and physical phenomena that, although not intrinsically stochastic, or at first sight so, are somehow reducible to this form? We know that Laplace's equation can be approximated to by a linear difference equation representing a random walk problem in which the probability that the particle will move from any grid point to any of the six neighboring grid points is the same, rendering the equation amenable to the Monte Carlo treatment; also that Fermi suggested long ago (as measured in "Monte Carlo era" time units) that this technique be applied to the wave equation, which is essentially a modified Laplace equation. But does every differential equation have to be linear if it is to submit to the Monte Carlo technique, etc.? The question we have posed is a sweeping one. The truth is, once again, that Monte Carlo is so young that any innovation is to be valued, even when it is not strictly new but merely "sees" already familiar matters in a fresh light.

A great virtue of Monte Carlo, apparent to its innovators, Von Neuman and Ulam, is that it provides a means for subduing complex problems (including some whose formal mathematical solution has been accomplished) that perplex us on the practical, application level because if traditional "hand" methods of computation are applied to them, numerical results are unconscionably slow in forthcoming. Monte Carlo is thus a scheme for

bringing the enormous power of the electronic computer to bear on problems. There are other such schemes. A given machine has certain potentialities for managing problems, these being determined by the available schemes for "laying out" problems and the engineering of the machine, and by the two in conjunction. The simpler of these schemes were probably in the mind of the machine's designer when he was diagramming its circuitry, but whenever a new scheme is invented, it may enhance the potentialities of existing machines as well as those yet to be constructed. Monte Carlo furnishes a grand strategy for attacking a certain species of problems, the processes built into the machine being the tactics for realizing the strategy. Viewed in this fashion, Monte Carlo is a more generalized form of coding, more powerful than orthodox computer coding because, in the sequence of devices leading from a problem's formulation to its practical solution, it comes earlier, and the "leverage" a device provides is roughly proportional to its priority of application.

Another virtue of Monte Carlo is its ability to pass in review before our eyes a vast variety of configurations emanating from a manifold process; configurations which, because of their diversity and numerousness, would take us years to experience in the real-life setting of the process. It is this property of the method that Professor Morse of M.I.T. values so highly for use in Operations Research. It is the amazing rapidity of the electronic computer that makes this a practical possibility.

To summon up the configurations, the process is analyzed into its elements, out of which the intrinsic ones, those from which the process can be reproduced without doing violence to its nature, are singled out for consideration. As a matter of practical computation, the number of these intrinsic elements whould not be excessive, since they join by combination to produce the configurations, and we know that in the combinatorial arithmetic applying to such situations, numbers mount very rapidly for small changes in the values of the controlling variables. Thus even if the number of distinct "forms" or "manifestations" that each element can assume is only two and there are n intrinsic elements, the resulting number of configurations is already 2^n (already 1024 when $n = 10$). Actually, each element is, as a rule, capable of many more "manifestations", often a continuous (infinite) array of them, and there is a frequency distribution specifying the probabilities with which they are assumed. In the usual case the mode of combination of the intrinsic elements to form the configurations is interdependent, so that these univariate distributions (strictly, their random variables) are not statistically independent, but conditional probability is always troublesome to work with, and as a practical measure we can overlook this dependence if it is not too large.

The procedure is then as follows: for each element we use an independent game of chance (a table of random numbers, if you will) based on the element's underlying distribution to pick the particular "manifestation" that is revealed at the moment, the different simultaneous "manifestations" being combined to give a particular configuration. By continuing to "spin" our roulette wheel or game of chance, the great variety of configurations will sooner or later come up, and this with the same relative frequencies that they would be generated by the process in real life. (Subject, of

course, to the approximations we have allowed -- the substitution of a small number of intrinsic elements for the totality and independent distributions for interdependent ones.) In practice we are embarrassed by the richness of configurations thrown up and an electronic computer is required to keep track of them. If some mode or average of the configurations is required, the computer can obtain it for us while "auditing" them.

The origins of operational gaming are distinct from those of Monte Carlo. Traditionally our military establishments, the Army and the Navy, have conducted war games not only to train their personnel in the handling of equipment and their own persons under circumstances more nearly resembling the conditions encountered in combat, but also to reexamine for the benefit of the general staff old methods of warfare and test and develop fresh tactics. It is the method of using a "material" model, scaled down several steps from the phenomenon it is used to represent, to enable the human mind to work out ideas that are too complex for it to retain. The architect uses it when he makes a plaster of paris model of the capitol or museum he is designing. On a more active level, that of design in motion, a football coach uses it when he puts his men through their paces in the field, evolving a new attack formation.

Although the writer did not consciously intent to develop the point, there is considerable identity of form and function between the football situation and the war games of the military. As he sees it, the most important aspect of operational gaming is this introduction of the factor of human psychology, particularly as it operates under conditions of stress such as competition, into a model that otherwise represents a purely mechanical or purely natural situation, i.e., a situation in which the human element is absent. If we are to rely on simulation devices, under which category operational gaming must be included, there seems to be no other way of introducing the human element than by the use of human participants. A theory of human behavior, especially in the area bearing heavily on the problem under study, could be employed in place of active, living human beings; but then whatever might be true of other levels, there would be no simulation on the human level.

The first of the papers on the program that the writer chairmanned had this property of combining mechanical means with the human element as provided by living beings. It might be better, in order to bring into relief the particular interplay of the human and mechanical factors in this paper ("The Differences in Experimental Data" by A. J. Eckles, III), reproduced elsewhere in these Proceedings, not to talk about it directly but to give the gist of a telephone conversation the writer had with its author.

The problem of measuring the effectiveness of a means of destruction, which for simpler weapons is reduced to that of calculating a hit probability, is an old one. If a new rifle was invented, or a new type of bullet, the "classical" method to ascertain its hit probability was to set up a stationary mount or screen and have it shot at from a firing line a fixed distance removed. The number of hits would then determine the hit probability.

Little or no attention was paid to the circumstance that several neighboring holes in the screen might represent the injury or death of the same soldier, and the problem of overkilling was thus largely neglected.

A more grievous error, fundamental in character, was to assume that an estimate of the rifle's damage capabilities under the static, unruffled conditions of a firing range could be equated to its power to damage on a battlefield. One could make theoretical allowances for the kaleidoscopic changeability of the battlefield and the impact on the infantryman's nerves of the bustle and fire, a sort of theoretical simulation, but might it not be more accurate to introduct these factors deliberately into the model, to a degree compatible with safety considerations, in the form of mobile human participants, moving target representations, etc.? Simple mechanical factors can continue to be corrected without simulation; e.g., one can qualitatively decide that a boat-tail bullet, which was superior to a flat-base projectile on the testing grounds because of the extra 1000 yards of range it gave, was nevertheless inferior in actual battle where the variety of obstacles reduces the importance of range, and it is incommodious to replace the rifle barrels that are constantly being worn out by the heavier bullet.

A first approximation, still quite crude, to the realism of the battlefield, suggested by this line of thought, is to substitute irregularly moving mounts for the stationary ones that are ordinarily used to determine the kill probabilities of simple weapons on a firing range. A target must enter one's visual field and be "centered" there before it can be fixed at accurately, and the adjustments that are necessary for a target popping at one suddenly are altogether different from those demanded by a stationary target at which one will be firing away for some time.

Eckles and his group at the ORO have been making more realistic determinations of the effectiveness of a tank, as determined by the training of its crew, the construction of its guns and turrets, etc., by having it ride down a trail and face "targets" that show up suddenly and then dart away. If these "targets" are "anti-tank guns" engaging in this limited war game according to certain rules, a conception of the effectiveness of a particular species of tank against anti-tank weapons is obtained. Still more realistically, one can have a tank platoon engage "enemy tanks" and "infantry" in a mock battle in the day or at night under given terrain conditions, the friendly platoon being assigned a specific objective to be taken with the assistance of a given quantity of aerial or artillery support.

Such simulated tank encounters have been used by others before. What distinguishes Eckles' efforts is the extreme lengths to which he has gone to achieve realism; the electronics laboratory at the ORO has wired the panels representing enemy tanks so that they light up to simulate opening fire, continue "firing" while they are intact, and burst into flames when damaged by armor-piercing rounds. One would imagine that the cost of conducting such an experiment, other than symbolically on an office checkerboard, is excessive, which it would be if one had to stage set it in the countryside from scratch; but the Army regularly conducts maneuvers that do

not differ immensely from this conception as part of its training program, and as Eckles points out, if engineers and scientists are willing to enter into a cooperative relationship with it, they can obtain a massive amount of information useful both to themselves and the military at little additional expense.

The other paper which will be commented on has to do with the application of Monte Carlo to compute lethal areas. "Lethal Area" is an old notion in ordnance research; it is that portion of an "initial" area in which an appropriate target will be incapacitated by a weapon system whose properties are known; the ratio of the two areas gives the probability that the target will be incapacitated when placed at random in the "initial", larger area, so that "lethal area" is properly a probabilitistic rather than purely analytic concept. This ratio is a kill probability with a geographic reference. Besides the area and the location of the weapon system in relation to it, there are many other parameters inherent in the system and the particular use to which it is being put at the time, all of them subject to probability distributions of their own, which determine the ratio.

We have already seen that the Monte Carlo method is able to evoke the myriad manifestations (configurations) of a phenomenon by playing a game of chance on each of the phenomenon's intrinsic elements. By using a table of random numbers to decide which value in its distribution of values any one parameter is to assume at a particular time, we can determine the form that the lethal area takes at the time; the aforementioned ratio is then determined automatically. We cannot go into the further details of Dr. Ehrenfeld's paper, which was classified, "confidential", since we desire to keep these remarks unclassified. A point in his favor is that there is provision for estimating the ratio of the lethal to the "initial" area by means of confidence intervals.

SOME DIFFERENCES IN EXPERIMENTAL DATA

A. J. Eckles, III
Operations Research Office

Perhaps the title of this presentation is somewhat a misnomer. But I do hope that it is not too misleading. Essentially, I would like to talk for a few minutes about some of the different types of experiments as I see them, and the necessarily different techniques of design, analysis and control which are required. I will not refer to technicalities such as the choice between a greco-latin square vs. a partial factorial, or whether we should use non-parametric or parametric techniques of analysis. In essence, these are only the tools of our trade, and should be adapted to the situation at hand. However, I might imply that the most suitable designs presently available for military field experimentation are the more simple ones, and the best techniques of analysis which meet the necessary assumptions (or lack of assumptions) in this type work are non-parametric.

I would first like to mention some relevant background material. The primary purpose of conducting military research is to provide us with data from which we can predict, with some degree of accuracy (upon which our lives, and perhaps even our freedom might depend), the outcome of future combat actions in which a variety of weapons systems are used. Once we can do this, it is then a relatively simple matter to select those systems which give us the highest probability of success.

Now we attempt prediction by a variety of devious means (short of actual combat) in which we construct models, extrapolate from performance characteristics, etc., until we finally reach conclusions and make recommendations as to the relative value of a particular weapons system.

But here we are faced with a major difficulty! Just what sort of performance data for each weapon system shall we use in our model? It is quite evident that if our models approach reality then they, too, will be affected by important changes in performance characteristics for the various weapons systems. We could, of course, ascribe a particular set of desirable characteristics to a new weapons system, and then determine the effects that such a system would probably have on the outcome of a particular type of battle. To a large degree this is done in the better grade Science Fiction novels, where we carry this extrapolation one step further (therefore becoming more realistic) and ascribe a particular set of characteristics to our human actors.

To be quite frank, I have been thinking of doing this as a preliminary step in the night-fighting program at ORO. But in this case, of course, I would prefer to dignify the process by giving it a different name than "Science Fiction"-probably just dropping the word "fiction" would help some. We could set up a particular battlefield situation in which the action takes place at night. Then we could examine the outcome of the battles if the opposing forces were variously equipped for night combat. For example, if the enemy had IR and we had white light; or the enemy had nothing and we had far IR imaging equipment; etc. After many machine hours and several volumes of reports, I could probably conclude that the better the performance characteristics of our fighting equipment and personnel combination, the higher would be our chances of winning a battle.

But we still haven't solved the problem of exactly which performance characteristics we should use in order to obtain a desired level of performance in the field or in actual combat.

When a new weapon is in the "drawing board" stage, the designers feel as though they have at least some idea of the future performance characteristics. We can say with some assurance, for example, that an automatic loading device in a tank will provide us with a higher potential cyclic rate of fire than manual loading; or that with a suitable rangefinder system we can obtain range data accurate enough to hit a man size target at 1,000 yards quite consistantly.

However, I'm proposing here that we can never hope to extrapolate from drawing board characteristics, manufacturer's specifications or even "Army Board" or "proving ground" type data and predict the relative effectiveness of a particular weapon in a combat situation. If we do this, we must be certain that we add the term "fiction" behind our endeavors in "Science" to avoid misleading our audience. In other words, we have at best hopelessly limited ourselves to a system of arm-chair philosophy because we choose to ignore the all important interactions between the so-called "human variable" and the weapon, and the higher order interactions between the man-machine weapons system and the conditions under which the actions take place.

Now I'm sure that it is not necessary to further justify to any of you the need for realistic experimental data upon which to base our predictions for the future. But the question I'm trying to bring out is: which of the many types of experimental data should be utilized in order to answer questions of importance to the Military?

I would like to present one example which will illustrate the nature of the problems we face. First, consider the selection of a rifle for combat. We can experimentally measure such factors as rates of fire, accuracy of the weapon when fired from a machine rest, barrel life, etc. These studies would not be what I would call Military field research. What the military is really interested in is the over-all casualty producing effectiveness of the man-machine system when various types of weapons are used. For example, the number of target hits (as different from the number of targets hit) is not a measure of a weapon's performance in the military situation unless we are willing to equate the killing of one man ten times with the killing of ten men one time each.

And while such factors as rates of fire and potential accuracy are undoubtedly related in some presently unknown and undoubtedly non-linear way to the combat effectiveness of the rifle-man weapons system, the only manner of actually predicting the effectiveness of such a system is to conduct a field study in which we use a suitable realistic criterion measure. And this is, I believe, at the present time the area of military research which presents the greatest problems: the development of realistic criterion measures which can be used in the conduct of field experiments.

It has often been said that in order to conduct "Field Experiments", the scientist moves his "laboratory" out into the "field" to collect his data. This is perhaps true in the non-military types of field studies

such as those currently being conducted in rocket research, lethal radius of burst from projectiles, barrel erosion, etc. But when we become involved in military research, which includes the utilization of military units with all the concomitant problems of man-machine interactions, and the host of differences attributed to the human variable, we must admit that the problems faced in field research are vastly different from those faced in the laboratory.

In the laboratory where we examine the relatively simple phenomenon (such as the fluttering of a relay, the time of projectile flight, growth of corn, behavior of rats, or the performance of memory), we can afford to indulge our whims and use complex experimental designs and their necessary techniques of analysis. However, in the area of military field research where the important problems are highly complex, we usually find that our requirements are much more efficiently met by quite simple designs, and even the simpler techniques of data analysis (primarily, of course, because these simpler techniques ((such as non-parametric statistics)) require that fewer assumptions be made about the conditions of data collection).

Now I appreciate your being patient with me as I may have wandered around the proverbial barn, but I felt that it was necessary to present some of the problems which have forced us to try a relatively new method of attacking the problems of military field research. In addition to the problems I've mentioned above (i.e., adequate control, suitable criterion measures, etc.), we also have the very practical problems of expense, both in money and in man-hours and equipment. We just have to face the fact that it is difficult to conduct the large number of field studies which are urgently required. (And here I would like to refer you to a talk tomorrow which will be made by Lt. Col. Clement, which will give many practical suggestions for urgently needed research.)

So in order to find a practical solution for obtaining realistic data, we are going to try and develope what we call a working symbiotic relationship between ORO Field Teams and Army Post-cycle training programs. We feel that at the present time there is a large source of data in the Army Training Programs which is going to waste simply because we have not yet developed suitable systems and techniques of data collection. By using such techniques there will be no shortage of experimental subjects, and our samples can be as large as we wish. The supplies and equipment available are, compared to previous field studies, inexhaustable. The only "real" cost to obtain this data is what is required for instrumentation and researcher salaries.

What we propose is truly a symbiotic relationship, not a parasitical one, for the military gain as much from these techniques as does the research worker, and in most cases even more. Their direct gains are primarily in the form of increased realism in the training program, and the concomitant increase in troop motivations.

In essence, this technique requires that we superimpose simple experimental designs and data collection techniques over the Army training

programs. Of course, the designs used must be simple to follow in the field to minimize the control problems and interference with necessary military proceedures. And in order that the resulting data have greater value, the instrumentation must not detract from the normal operations, but rather increase the realism where possible.

Before spending a few minutes describing one application of this symbiotic relationship, I would like to discuss some of the differences between data obtained in this manner and data which might be obtained from the conduct of a specific experiment. Essentially, we would find the following differences.

1. Our control is not always what we would like. In many cases we are in the position of astronomers who can only record the events as they happen, but are limited in the manipulations which they can perform. In other cases safety precautions force us to utilize situations which are unrealistic.

2. In compensation for our lack of rigid controls, however, we are able to utilize continuing cycles of training, thus increasing our sample size far beyond what we could expect to demand in a specifically conducted experiment.

3. We have time between runs to "Debug" our program, improve our data collection system, and build our design as we progress. (Though this might violate some of our current thinking; i.e., that we complete our experimental design, including the methods of data analysis, prior to the conduct of the study.)

I would now like to spend a few moments in giving you a brief description of how we plan to utilize this technique of "Symbion" in order to collect one type of experimental data.

Fort Stewart, Georgia, is presently conducting as part of their regularly scheduled training program a problem which involved a tank platoon in a night attack, using live ammunition. This problem was called the T-2 exercise. Essentially this was a free-play exercise in which the platoon leader was assigned the mission of taking his objective by a night attack, when the objective was defended by enemy tanks and infantry. In this attack he was supported by a 60-inch searchlight. The enemy tanks were represented by the standard 6x6 panel targets, and the enemy infantry by the standard Type E targets. The attacking platoon would be notified by radio that they were under enemy fire at an appropriate time during their advance, and they would then undertake to fire upon the targets until all of their ammunition was expended.

It was the normal conduct of this T-2 exercise and the close cooperation by the officers and men of Fort Stewart which have made it possible for the ORO field team to design and conduct the present research project in night fighting. On the part of Fort Stewart, they have permitted the use of their training program, with the necessary modification, to change the T-2 exercise into a veritable "laboratory-in-the-field." This has,

of course, required additional effort from both the officers and supporting personnel, and a willingness to put up with the needs and desires of the scientist. But in return for these additional burdens, the scientists from ORO have added realism and meaningfulness to the training program.

For example, the Electronics Laboratory at ORO has designed and supplied a new type target to simulate the enemy tanks. These targets, rather than being simple, passive panels, initiate the engagement by simulating opening fire upon the attacking platoon. The targets then continue to "fire" upon the platoon being tested until they are hit by an AP round (small arms fire and small fragment hits have no effect). When finally hit by an AP round, the newly developed ORO targets stop firing and burst into flames to simulate a burning enemy tank.

Throughout this rather realistic engagement, the field team from ORO is busily collecting and recording appropriate data which will provide a measure of the platoon's effectiveness in night combat.

Over a period of several months, by testing a number of units equipped with a variety of night fighting equipment - such as tank mounted fighting lights, infra-red equipment, pyrotechnics, etc. - this joint ORO-Fort Stewart project will not only better prepare these units for night combat, but also provide us with the answers to a number of questions about our present capabilities for night operations. Questions such as the relative fire effectiveness of armored platoons when equipped with various types of equipment, hit probabilities, and rates of fire of our tanks under various types of illumination, etc., will be at least partially answered by the first phase of Project SYMBION.

In summary, then, I've been making a plea for more data of the type which is obtained from operationally realistic field experiments, in contrast to the type of data obtained in most "laboratory-type" or "proving-ground-type" studies. And I have proposed a possible technique, "SYMBION", for obtaining this type of data with minimum expense. In fact, ORO has designed such a program with the cooperation of the Officers of Fort Stewart, Georgia, which will begin this October (1956).

THE APPLICATION OF DESIGN OF EXPERIMENTS AND MODELING TECHNIQUES TO COMPLEX WEAPONS SYSTEMS

E. Biser and M. Meyerson
Signal Corps Engineering Laboratories

1. Purpose. The purpose of this paper is to outline a conceptual plan and framework that was used to establish a Design of Experiments for a weapons system. Further, the paper will indicate the application of a model for analyzing the system.

2. Background. During World War II, it became apparent to antiaircraft experts that, although individual antiaircraft gun batteries were relatively effective against single targets, the defense of a critical objective, as a whole, against large target raids, was relatively ineffective. Consequently, military requirements were formulated for an integrated system, wherein all the processes of AA defense could be coordinated, resulting in an overall increased system effectiveness. A system was proposed by the Signal Corps Engineering Laboratories, Fort Monmouth, New Jersey, approved, developed, constructed, installed and readied for test. This paper described the processes which were involved in developing the test plan, some of the general tests, and the final consideration of the efficacy of this complex system. Although the system has been completely tested, broken down to basic sub-systems and given to other agencies for research and development, it has served this purpose well, and the concepts described herein have formed the basis for evaluating all other systems of this type, under Army Signal Corps cognizance.

3. Discussion.

a. Design of Experiments. Although many definitions exist for this term, a most appropriate one for the purpose of this paper might be that depicted in Figure 1.* Here the system is shown as a series of symbols depicting the man-machine combinations and interactions, all combining to produce a desired objective. The purpose of the experimental design, then, is to adequately define the desired objective (or objectives), test the system to measure that objective, and then to determine the contribution of each system block toward the desired objective.

In the light of the basic objective, we were confronted with the fact that we had a new system that would obviously be compared with an existing system prior to the time Army Staff might accept it for standardized issue. Hence, we considered it advisable to analyze and to clarify the following semantical equation: our goal is to measure the improvement of this newly-proposed Weapons System over existing Antiaircraft Defense Systems. The sentence can best be investigated by symbolizing "Improvement" by (1); "Newly Proposed Weapons System" by (2), and "Existing Systems" by (3), as follows:

(1) Improvement: The following relevant questions naturally present themselves concerning the concept of improvement:

* Figures appear at end of the article.

(a) What is meant by improvement?
(b) What are its criteria?
(c) What magnitude of improvement is to be discussed and analyzed?
(d) What is the optimum method of measurement of improvement?
(e) Who has to be convinced that the method of analysis and especially that the design of experiment has yielded significant and worthwhile results regarding improvement:

1. What is meant by Improvement? There are two main areas where improvement is urgently needed as follows:

a. Rational distribution of fire. By this is meant a firing doctrine or rationale that optimizes minimum damage to the defended area, attrition or prevention of penetration by spreading AA fire over the entire attacking raid.

b. Improved intelligence on air raids, i.e., with respect to detection and identification of targets. While rational distribution of fire is readily given to quantitative evaluation, improved intelligence, although contributing greatly towards overall system effectiveness, is not easily quantifiable. It should be noted that it may not be possible to evaluate the measure of rational distribution of fire without taking cognizance of improved intelligence on air raids.

Here, the test designer quantifies the basic test objectives, for which all following concepts and the actual tests will be designed.

2. What is to be the criterion or criteria of Improvement? This is a vital question since it will have a great bearing on the type of defense index to be quantified. The criterion of improvement may consist of the optimization of defense per dollar spent. This concept can be further narrowed down and particularized to the following quantifiable parameters:

a. Least damage to the defended area per dollar spent on antiaircraft defense for that area. This indicates that the aim of building a defense system is to prevent damage (i.e. physical, psychological, productive, et al) to a defended area above a predetermined minimum. Here damage is the independent variable and is established at a value above which the war potential of the area is seriously or completely hampered.

b. Maximum damage to enemy raiders per dollar spent on antiaircraft defense. This stresses that the objective of building a defense system is to insure a predicted maximum attrition (i.e. the loss to the enemy of his attacking aircraft and consequent destructive potential) for a given area. Here attrition is the independent variable and is established at a value above which a certain number of potentially destructive enemy aircraft would elude the defenses.

c. Lowest probability of penetration by enemy raiders into the defended area per dollar spent on antiaircraft defense. This states that the goal of building a defense system is to insure the prevention of a certain percentage of enemy penetration to a defended area. Here prevention of penetration is the independent variable and is established at a value below which a certain number of potentially destructive enemy aircraft would penetrate the defenses.

Here the designer offers some food for thought for which the objective may be measured.

3. Magnitude of Improvement: It is necessary to assign a measure-number to the concept of improvement, since this number will tend to give a decisive indication of the efficacy of this integrated defense system. It is estimated that the following magnitudes of improvement of the newly-proposed system over existing systems might be expected:

a. For low kill probability weapons in the system subjected to saturated types of raids, a small improvement might be expected with respect to the three parameters mentioned above, since even coordination of low kill-probability weapons does not materially increase their overall effectiveness (determined by allied studies). The contribution towards this overall improvement is due to rational distribution of fire, as well as to improved intelligence on air raids.

In the case of these low kill probability weapons, however, because of their low kill probability, improved air raid intelligence, though not rigorously quantifiable, appears to contribute most to overall improvement with respect to the aforementioned three parameters. In the light of these considerations, it would appear that experimental research could better be concentrated on improvement of intelligence, and analytical research pursued in the area of rational distribution of fire for these weapons.

Here the designer actually recommends where tests and analysis could best be utilized for maximum economy.

b. For other weapons, because of their higher kill probability, it is anticipated that a greater improvement with respect to the aforementioned parameters could be attained. In this case, for reasons alluded to previously, it would appear that experimental and analytical research should be equally apportioned with respect to rational distribution of fire and improved intelligence.

Here, again, the designer indicates the type of effort to be expended, but for different weapons.

4. Optimum Method of Measurement: The consideration of optimum (but practical) methods of measurements and comparison of the newly-proposed system with existing systems entail the following two modes of comparison:

a. Comparison on a simulated basis, with only the output (i.e., weapon battery firing) being simulated. This means that air-craft will actually be flying and effective kills calculated on a simulated weapon battery firing basis.

b. Comparison of systems by simulating both the input (Target Simulator) with aircraft not flying, and output, (AADECAR-Antiaircraft Defense Effectiveness Computer and Recorder) with weapon batteries not firing.

Here the designer specifies the nature of the test and even some of the major test equipment to be used.

5. Personnel Interested in Analysis and Findings: Three different primary agencies and interests are concerned with the results of the analysis and the findings of the experimental design:

a. Army Antiaircraft Command, the ultimate user of the equipment, is interested from the standpoint of operability, reliability, and overall effectiveness, as a tactical weapons system.

b. Continental Army Command, as the experimental arm of the Army for systems of this type, is concerned with the verification of operational concepts set forth in the military characteristics, as well as with operability and reliability.

c. Signal Corps, as the technical service, is concerned with obtaining technical data on all significant factors which affect the overall system design.

Here the designer has indicated that the final test results must be in such a form as to be readily understandable to different agencies, with different interests, all of whom will draw conclusions regarding the system efficacy.

(2) The Newly Proposed System. Since the system is not a static model, it is worth noting that a description of that system falls into two categories as follows:

(a) The present installation consists basically of detection, identification, data processing, tactical evaluation, assignment, acquisition, tracking and engagement functions, with interconnecting communications (further details will not be revealed here because of the classification of the information, and since it is not particularly germane to this discussion).

Here the designer actually described the system, so that the establishment of the mathematical model, and the ultimate conclusions regarding the contribution of each of the major system blocks will have the same meaning.

(b) A short term improvement version of the present installation with improved technical, tactical and operational facilities (again no further details are necessary here).

Here, again, the designer recognizes the logical progression to a slightly improved model which will also be covered by this evaluation.

(3) Existing Systems: The existing system is used as a reference system with respect to which improvements are to be measured. This system is then defined (not in this paper) in the same manner as was the newly proposed system.

With the foregoing clearly established, the designer would then focus attention on some of the crucial factors that are likely to affect system effectiveness. Some of the following factors, singly and severally, were considered as follows:

Broad Factors:

Performance of man-machine system subjected to saturated types of raids.
Performance of man-machines under conditions of jamming and clutter.
Performance of data processing equipment in response to diverse and complex courses.
Capability of human operator to perform assigned tasks under adverse conditions of complex and saturated raids.

Detailed Factors:

Rate of entry of targets.
Reliability and resolution of identification sets in the system.
Effect of radar resolution at ranges of primary interest on the operation of the system.
Resolution and readability of displays and boards.
Effect of battery acquisition time on system effectiveness.

Having thus established the conceptual framework for the test, the next step was to model the system so that it might best by analyzed.

b. The Modeling Approach.

Although no stranger to science, no term is more frequently used in current literature on operations research than that of model. Indeed, the concept of model has come to connote the hallmark and canon of scientific method and intelligibility. Scientists have given substance to the ideas embodied in their theories by means of mental pictures or physical models, such as models of ships, railroads and airplanes (to study flight characteristics), just to mention a few static models.

The questions naturally arise: what is a mathematical model, and how is it helpful in describing the functions of a large scale man-machine

system? what is it purported to do? what are its constitutive elements? how is it constructed, etc.?

One word singularly expresses the most essential meaning and significance of model:it is the term symbol. A symbol is a representation of an event. This term, then, is the key to the following compact definition of a mathematical model. A mathematical model is a symbolic representation of a system (the domain of phenomena under investigation).

(1) Weapon System:

Before analyzing the structure of a model, let us pause briefly to review the peculiar nature of a weapon system. A weapon system is an organization of men and equipment designed for operation and use against a class of entities known as targets. In order to carry out its overall function, it must also carry out many complex subfunctions. The function of the system can be functionally subdivided into many different activities, depending upon the kind and types of activity to be carried out. Each functional activity requires certain quantitative inputs to be converted by this functional activity into another quantity called output.

A Weapon System, for instance, consists of observation units, information processing units, and action units. It contains communication facilities to handle classes of information such as weapon information, target information, etc.

The concept of Model is predicated on the assumption that it is possible to abstract, from a complex system, certain persistent and discernible relationships and to mathematize and quantify these relations with a view of describing the behavior of the system. The initial stages of modeling consist of devising concepts that describe the purpose, functions, operations, pertinent parameters or state variables, all of which go toward erecting the frame of reference for the mathematical model to be operative. This was accomplished in the earlier portion of this paper. The goal is to construct a model so that, by studying its characteristics, it will be possible to deduce the state of the system (the output of the system) under varying conditions (raid configurations).

(2) The Objective:

The main objective is to construct a theoretical-experimental model, hereafter to be referred to as a mathematical model for evaluating the efficacy of the weapon system and to evolve intrinsic and comparative criteria and measures of effectiveness. The aim of the model is to establish a theoretical-experimental structure within which the large scale man-machine system is to be evaluated with respect to certain predetermined criteria of effectiveness, such as maximum defense, maximum attrition, etc. The point of departure is that the best way of describing and evaluating the large scale system is to construct a model involving quantifiable parameters to predict the dependence and variation of each

pertinent parameter on each functional activity of the system. The model should exhibit how the various functions of the system, such as detection, identification, data processing, tactical evaluation, assignment to weapons, acquisition, tracking, engagement and weapon characteristics affect one another, i.e., how they are interrelated and interconnected.

The model envisaged here is not an aprioristic one, namely, one totally divorced from test data and superimposed on the system, without recourse to test data. It is not an axiomatic model so characteristic of abstract mathematical systems defined implicitly by a set of axioms without regard to any significance and meaning attributed to the symbols used. (The significance of the symbols, in an axiomatic model, is governed solely by the linguistic rules laid down by the axioms.) The model to be operational in the experimental sense is not to be construed as a mathematical scheme, or as an ensemble of apriori concepts to be arbitrarily imposed on the operations of the system.

Such concepts untested and not subjected to experimental control would be sheer intellectual ghosts without operational efficacy and meaning. It is clear that the importance of test data cannot be gainsaid. Nor can they be dispensed with in the modeling approach. It is a realistic system (or a class of structurally similar systems) whose behavior, output and time response are to be described and predicted by a theoretical model. The weapon tests (with live and simulated inputs) will provide data that, when properly reduced, will provide unbiased statistical estimates of significant parameters. It is these parameters that are to form the basic structural elements of the model.

The test data will provide the quantitative empirical data to fill out the model and to validate the model experimentally. It is the model, through its predictive efficacy, that is to describe and to predict the response of the system to varying inputs (raid configurations).

(The flow chart in Fig. 2 profiles, by block diagram, the distinctive, logical, and sequential steps involved in system modeling.)

(3) The Mathematical Model: The Weapon System is functionally divided into the following activities or units:

Detection, identification, data processing (manual and/or automatic), tactical evaluation, assignment of weapons to target, acquisition, tracking, firing and ultimate kill. The partitioning of the overall function of the system into these subfunctions was made advisedly consonant with the concept of a weapon-complex as a dynamic or a time-response system. It was natural to undertake measurements of time intervals (time delays) corresponding to these functional activities. These time delays are to be described as mathematical functions of input parameters, such as range, radar cross section, velocity, etc.

The model is structurally isomorphic to a logical syllogiam in which the system is the major premise, the input is the minor premise and the output is the conclusion. This basic syllogiatic description of a system has important implications. The conceptual scheme:

input - system - output

can be expressed as follows:

Given a system and a class of inputs to determine the output response characteristics of the systems. It can be formally expressed in the following symbolic equation:

$$X_o\ (t) = \int_{t_o}^{t} S\ (t)\ op\ X_i\ (t)\ \Delta\ t$$

this is symbolically analogous to an integral equation.

X_o (t) ≡ the set of output responses of the system (to be described subsequently)

X_i (t) ≡ the set of inputs to the systems.

S (t) ≡ the set of transfer functions characterizing the system.

op. ≡ the coupling of the inputs to the system.

With this in mind, the model is to consist of the following structural units:

(a) Model Parameters: These consist of elevan (11) functionally defined time delays T_1 to T_{11}. In fact, these parameters are probability distributions of the time intervals associated with various functions of the systems. It is to be noted that the term "parameters" is not to be construed as a statistic such as mean, variance, etc., but as functional variables which are in turn to be related to input variables.

The boundaries of the time intervals, t's, are functionally defined as follows: *

Functional Definition

t_1 ≡ time target entered system (detection and identification) time[5] of the beginning of telling the first early warning plot for a new target from higher headquarters, (or the facility simulating it) to this system. Recorded on magnetic voice tape.

* Superscripts are explained on Page ; remaining symbols are explained on Page .

Functional Definition

t_2 = time target entered a track-while-scan computer. (data processing)	time[5] of first variance in track-while scan computer IBM output.
t_3 = time when track-while-scan computer first establishes a smooth track. (data processing)	1) time[5] of first x punch for each target assignment to a track-while-scan computer. 2) time[5] of appearance of white light next to channel number at the left of the Engagement Status Board.
t_4 = time of first height information received by track-while-scan computer from height-finding radar. (data processing)	1) time[5] of height dots on tactical display and background height report to confirm that height was not entered from early warning information. 2) time[5] of first variance in track-while-scan computer recording output. Punch out and background height report to confirm that height was not entered from early warning information.
t_5 = time target was assigned to a battery. (assignment)	1) time[5] of appearance of battery letter on tactical display. 2) time[5] of "on" signal in battery recording for the first time for a target-battery combination.
t_6 = time target first examined at battery. (assignment)	time[1] of first "on" punch for each target-battery combination.
t_7 = time of target designation to battery tracking radar. (acquisition)	time[1] of first "on" punch for each target-battery combination.
t_8 = time of target lock on by battery tracking radar. (tracking)	time[1] of first "on" punch for each target-battery combination.
t_9 = time of fire (engagement)	1) time[1] of first "on" punch for each target-battery combination.

2) time[5] of the appearance of red firing light on Engagement Status Board for each target-battery combination.

t_{10} ≡ time of missile impact (engagement) — time[1] of first "on" punch for each target-battery combination.

t_{11} ≡ time of battery "ready" for next assignment. (transfer time) — time[1] of first "on" punch for each target-battery combination.

t_{12} ≡ time the next target is designated. (transfer time) — time[1] of first punch for a target-battery combination which is preceded by punches in any column referring to another combination.

The superscript 1 and 5 indicate time measurements with one and five seconds accuracy. The time intervals T_i (i = 1 to 11) are accordingly defined as follows:

(1) * T_1; time of entry of each target into system from early warning information to time each target is entered into a track-while-scan computer ($t_2 - t_1$) (detection and identification).

(2) * T_2; time each target is entered into a track-while-scan computer to time of first smooth narrow gate tracking ($t_3 - t_2$) (data processing).

(3) * T_3; time of first smooth narrow gate tracking to time height information is first available from height finder for each target ($t_4 - t_3$), (data processing).

(4) * T_4; time height information is first available from height finder to time target is assigned to a battery for each target and for any one target, each battery ($t_5 - t_4$) (tactical evaluation).

(5) T_5; time a target is assigned to a battery to time target is first examined on Battery Commander's PPI for each target combination battery ($t_6 - t_5$) (assignment).

(6) T_6; time target is first examined to time target is designated to tracking radar for each target-battery combination ($t_7 - t_6$) (acquisition).

(7) T_7; time target is assigned to tracking radar to time tracking radar locks-on target for each target-battery combination ($t_8 - t_7$) (acquisition).

(8) T_8; time tracking radar locks-on target to time missile is "fired" for each target-battery combination ($t_9 - t_8$) (tracking).

(9) T_9; time missile is "fired" to time of missile "impact" on target for each target-battery combination ($t_{10} - t_9$) (engagement).

(10) T_{10}; time of missile "impact" on target to time battery is ready for reassignment for each target-battery combination ($t_{11} - t_{10}$) (transfer time).

(11) T_{11}; time battery is ready for reassignment to time a new target is designated to that battery for each battery and for each new target ($t_{12} - t_{11}$) (transfer time).

* Currently the system instrumentation does not permit explicit separation of T_1 and T_2. If they cannot be separated implicitly, or through a minor instrumentation change, they will be carried in the analysis as $T_1 + T_2$. The same applies to T_3 and T_4.

MODEL PARAMETER	INSTRUMENTATION
T_1	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number, track-while-scan computer number, EW (early warning) voice and plots.
T_2	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number and track-while-scan computer number.
T_3	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number, track-while-scan computer number, early warning height, track-while-scan output (x, y, h, $\dot{x}$, and $\dot{y}$)
T_4	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number, track-while-scan computer number, battery number, order of assignments to batteries, correlation time for each target, track-while-scan computer output, Command and Status signals.
T_5	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number, track-while-scan computer number, battery number, whether this is the 1st, 2nd, etc. target handled by a given battery, track-while-scan computer output, Command and Status signals.
T_6	a_1, a_2, R_r. c, n, R_s of targets. Also AF target number, track-while-scan computer number, battery number, whether this is the 1st, 2nd, etc. target handled by a given battery, track-while-scan

MODEL PARAMETER	INSTRUMENTATION
	computer output, any track data by other batteries on target now being assigned to a battery radar during T_6, Command and Status signals.
T_7	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number, track-while-scan computer number, battery number, whether this is the 1st, 2nd, etc. target handled by a given battery, track-while-scan computer output, any track data by other batteries on target now being assigned to a battery radar during T, Command and Status signals.
T_8	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number, track-while-scan computer (target) number, battery number, whether this is the 1st, 2nd, etc. target handled by a given battery, track-while-scan computer output, battery track, Command and Status signals.
T_9	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number, track-while-scan computer (target) number, battery number, whether this is the 1st, 2nd, etc. target handled by a given battery, track-while-scan computer output, battery track.
T_{10}	a_1, a_2, R_r, c, n, R_s of targets. Also AF target number, track-while-scan computer (target* number, battery number, whether this is the 1st, 2nd, etc. target handled by a given battery, track-while-scan computer output, battery track.
T_{11}	a_1, a_2, R_r, c, n, R_s, ϕ of consecutive targets handled by a given battery. Also AF target numbers, track-while-scan computer numbers, battery number, track-while-scan computer output.

where

$\dot{x}$, $\dot{y}$ = target velocity components
x, y, h = target position coordinates
a_1 = aspect to battery
a_2 = aspect to operations center
ϕ = angle between radius vectors from a battery to consecutively handled targets (plan view).
c = concentration
n = number of targets
R_r = range of resolution
R = slant range of target at initial point of time interval.

(b) Input Variables: The input variables constitute those characteristics and features of raids that go to determine wholly or in part the effectiveness of an AA defense system. The input variables include height, velocity, radar cross-section, early warning information, resolution range, path of the target, etc. It is to be noted that the mathematical relation of the input variables to each of the model parameters, the delay intervals T_1, T_{11}, is of paramount importance to the creation of the model.

(c) System Configuration Parameters: These include the number of batteries, their location and relative distance among them, and the number of operating batteries. Although these parameters primarily refer to the geometry of the system, they include weapon characteristics such as kill probability curves, maximum and minimum firing ranges, etc.

(d) System Logic: The system logic essentially describes how the system operates on input data, what the operators do, how assignments are made, under what conditions open fire commences, etc. The system logic thus refers to the operating procedures with respect to a fixed set of input variables; it contains standard operating procedures, as well as assignment doctrines.

(e) Measures of Effectiveness: These constitute criteria that give an explicit measure of the extent to which the defense system is attaining its main objective. The concept of measure of effectiveness, in effect, implies that the goal and operations of the system are clearly, significantly and explicitly stated. In fact the model in its entirety is built around the measures of effectiveness which, in essence, define the goal of the system. These objectives, as defined by the criteria of effectiveness, must be self-consistent, since it is impossible to make consistent fundamentally inconsistent goals.

An index is a number, a measure-number, and this number can hardly be conceived without criteria by which the effectiveness of the weapon system is to be assessed. An index is a measure-number indicative of the effectiveness of the system with respect to predetermined criteria of effectiveness. A Defense Index is the selected criterion quantified to measure the output of the defense system. Thus, there is no index without selected criteria and without operational data (test data). For example, if maximum attrition (the maximization of the expected number of targets destroyed) is the criterion chosen, the system subjected to a given class of raid, may have an index of 0.3 with respect to this criterion.

It is evident that the primary objective of a defense system is to "score" against enemy planes. Hence, it follows that the statistical distribution of planes destroyed by the system would yield all the information needed to assess the capability and efficacy of the system against enemy targets. Such a distribution will contain the expected number of targets destroyed (E), the probability of non-penetration (P_{np}), i.e., the probability that all the planes in the raid are destroyed, the probability that, at most, a specified number of planes survive, etc.

The mathematical model envisaged cannot be committed exclusively to the criteria of maximizing P_{np} or E, the concepts of maximum defense and maximum attrition respectively. In fact, it is desirable to devise a more general class of criteria, (in view of the advisability of considering all possible enemy strategies containing maximum attrition (E) and P_{np} as "limiting" criteria. Without unduly belaboring the point, it is worth noting that realistic situations may change to the extent of requiring the maximization of expected number destroyed and, under varying conditions, the maximization of P_{np}, especially if the damage to the defended area is catastrophic if one or several planes penetrate the defenses.

In short, given a kill probability density function, a damage function of the number of enemy targets penetrating the defended area, it cannot be stated categorically that damage to the defended area will be minimized by maximizing either E or P_{np}. In order to minimize the expected damage to a defended area, the entire distribution of the number of planes surviving (or destroyed) and not merely P_{np} or E (T), the expected number of targets destroyed, needs to be superimposed on the appropriate damage function.

To return to the conceptual scheme:

input - system - output

We can see that equipment, system logic, and system configuration constitute the system and its operation. The output is given in terms of the multinomial distribution (P_i), the probability that exactly i targets are destroyed in a raid, i = o, ...n, where n is the number of attacking aircraft; $\{P_i\} \equiv (P_o, ..P_i, ...P_n)$.

$P_n = P_{np} \equiv$ the probability of non-penetration.

This distribution, together with the criteria of effectiveness, will determine the desired output (with respect to the selected criterion).

To summarize: The mathematical model consists of the (1) model parameters, (2) the input variables, (3) system configurations, (4) system logic, (5) and measures of effectiveness.

c. The Monte Carlo Technique:

With the multinomial distribution, P_i, as the primary output of the Model, the question naturally arises how this distribution is calculated. It would indeed be desirable to determine analytically the exact distribution of the planes destroyed. At this stage, however, this goal is well night impossible of attainment. It should be noted that this Model is a probabilistic one because of the stochastic nature of the model parameters and weapon characteristics. The Monte Carlo method is eminently suited to estimate the distribution, since this method is essentially a sampling experiment making use of large tables of random numbers. The term "Monte Carlo" is descriptive of a whole class of

calculational techniques called stochastic because of the use of random numbers.

The aim of this technique is to find a stochastic process that has a distribution corresponding to the physical situation under investigation. (Strictly speaking, this method cannot yield the entire distribution.)

A high speed digital computer is utilized to implement the substitution of a stochastic procedure for an analytic model of the system. What the computer is actually doing is to sample from the exact distribution in order to estimate it. The exact distribution of the real situation is approached more and more closely as more runs are made on the computer (this is based on the law of large numbers). Representative samples are being followed through their histories to obtain an approximation to the entire distribution.

The computer samples in a random manner from each of eleven time-delay (T_i) distributions: For a given raid configuration consisting, say, of n aircraft, one value of each T_i is obtained for each target. Thus corresponding to target A_1, T_1 up to T_{11} are obtained (some T's may be zero if the target fails to transit the entire system).

$$A_1 \; : \; T_1 \; \ldots \; T_{11}$$

$$A_n \; : \; T_1 \; \ldots \; T_{11}$$

A given raid will be rerun between 50 to 100 times. These runs will produce samples of T_1 to T_{11} (inclusive). The factorial design consists of 96 blocks with four replications in each block making a total of 384 data.

The faster the computer program, the larger the sample size, the narrower will be confidence intervals for the estimated distribution parameters. The sampling distribution is multinomial with the following parameters: $P_o \; P_1 \; \ldots, P_n$, where P_i is the probability that exactly i planes are destroyed, n is the number of planes in the raid. A finite number of raids will be selected to facilitate the correspondence of the response surface of the system (in terms of kill probabilities) to the multidimensional space of input variables.

(4) The Nature and Efficacy of the Model:

This is a stochastic (probabilistic) model, since distributions of the model parameters (the time delays) are involved. Corresponding to a sample of size N, of each T_i, there is a regression equation:

$$T_1 = T_i \; (v, \; h, \; R_s, \; c, \; n, \; R_r, \; a)$$

The parameters v, h, etc. are randomly varied in order to obtain samples. T_1, say, may be given as:

$$T_1 = .573 \; v + .672 \; h + 5.734$$

The mean of each T_1 is estimated for given values of the parameters.

Corresponding to each raid sample of size N, there will result frequencies for P_o, P_1, ... P_n. These constitute the response surface (with the aid of interpolation) of the system, with respect to a raid of a specific type.

A raid configuration is characterized by the input values of (v, h, R_s, c, n, R_r, a). A set of values, one for each parameter, (v, h, R_s, c, n, R_r, a) is defined as a raid vector. This is the input vector.

There are N samples of each input vector and consequently, at most, N values of P_o, P_1, ... P_n for each vector. Thus, there is a one-to-one mapping from each raid (R_j) to its corresponding response surface:

$$R_j \longrightarrow (P_i)_j$$

or: $$(v, h, R_s, c, n, R_r, a)_j \longrightarrow (P_o, P_1, \ldots P_n)_j$$

where: $$R_j \equiv (v, h, R_s, c, n, R_r, a)_j$$

The ultimate goal is to find explicit expressions of the output response surface (P_o, P_1 ... P_n) in terms of specific raid inputs where:

$$P_i = P_i\,(v, h, R_s, c, n, R_r, a)$$

The aim is to obtain a regression equation of kill probabilities in terms of height, velocity, range, number of targets in a raid, etc. This is possible only if a class of admissible raids is treated as one ensemble.

4. Summary and Conclusions:

a. The initial design of experiments established the conceptual framework around which the model was derived and the test designed. It stated the test objectives, the test criteria and the generally anticipated results.

b. A stochastic (probabilistic) model of a weapon system was constructed to describe and predict the (time response) output characteristics of the system for given inputs (raid configurations). This was accomplished by partitioning the overall functions of the system into subfunctions and their corresponding time delays, the model parameters, and finding mathematical relation of each time delay in terms of input parameters (the raid characteristics). The model parameters are estimated by the Monte Carlo method which is essentially a combination of numerical analysis and sampling theory.

The model contains a fixed physical system, and assignment procedure, weapon characteristics, and a standard operating procedure. The primary output of the model is the probability distribution of targets destroyed for a class of admissible raids. This distribution yields the

probability that exactly $\underline{i}$ targets out of a raid of $\underline{n}$ attacking targets are destroyed for each $i \leq n$. All possible effectiveness criteria are expressed in terms of the primary output.

The model contains a flow chart of a computer program which can be coded for any computing machine, so that, given the characteristics of a given raid and an assignment procedure, the corresponding system response, in terms of kill probability, can be computed.

The model is flexible: as the system configuration parameters change, the distributions of the time intervals change accordingly. It is thus possible to gain insight into ways of improving the system. It is to be noted that these parameters include kill probability curves. The model will make it possible to predict the behavior of the system when one (or several) of the time intervals are increased or decreased.

5. Acknowledgements:

We take this opportunity of acknowledging indebtedness to the Operations Research Group of the University of Michigan, for the work done on the conceptual and instrumental aspects of the model; and to the personnel of SCEL Field Station No. 4 (9584 TU) at Fort George G. Meade, Maryland, for their contribution to the instrumentation of the tests.

APPLICATION OF "DESIGN OF EXPERIMENTS" TO "A SYSTEM"

DESIGN OF EXPERIMENTS "A SYSTEM TEST" — THE DETERMINATION IN ADVANCE OF THE DESIRED RESULTS OF TESTS, IN ORDER THAT A PLAN MAY BE ESTABLISHED WITH THE AID OF STATISTICAL THEORY, FOR THE TESTS NECESSARY TO ACHIEVE THOSE RESULTS IN THE SHORTEST TIME, AT THE LEAST COST, FOR ANY DESIRED DEGREE OF CERTAINTY.

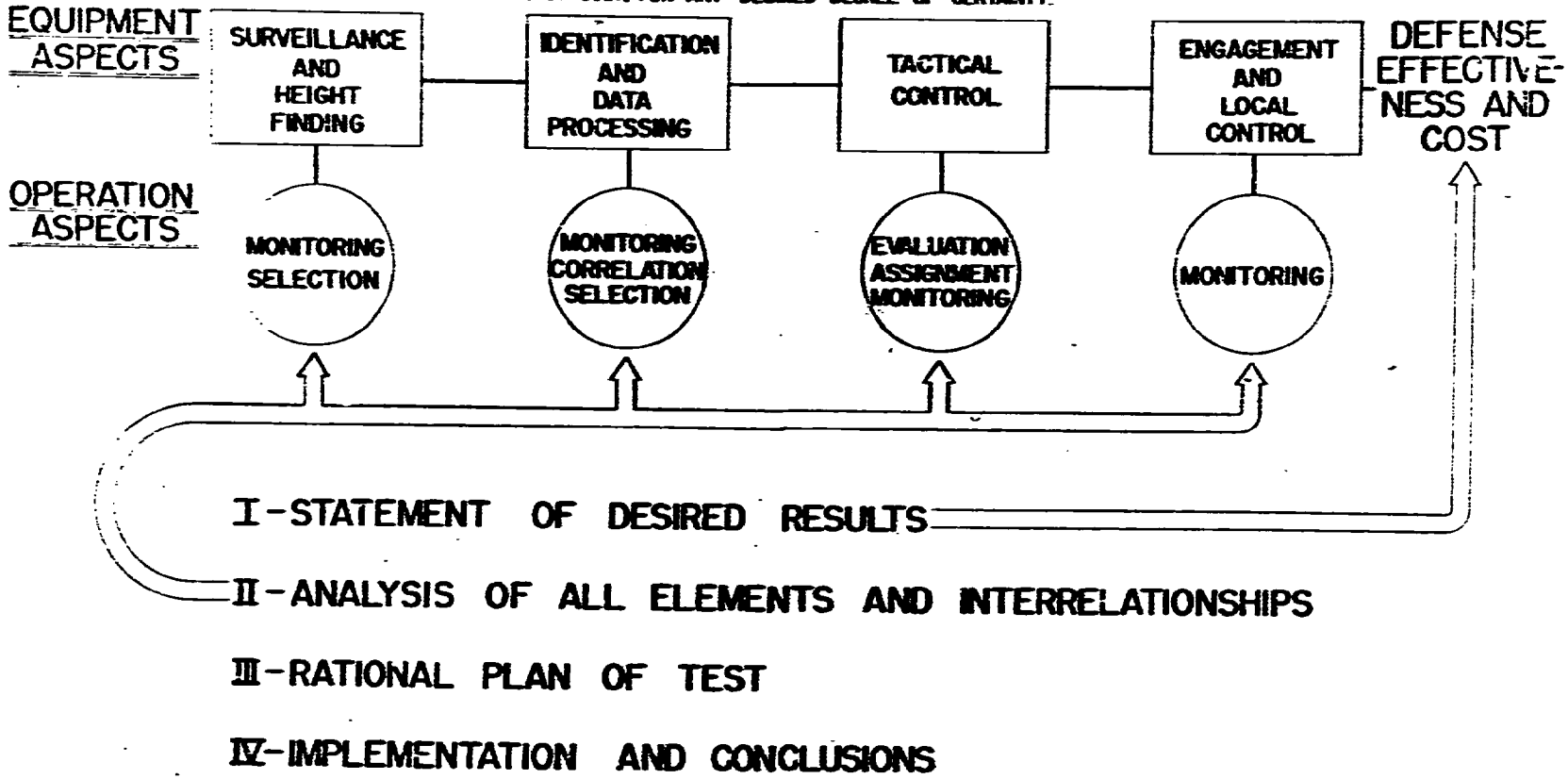


FIGURE 1.

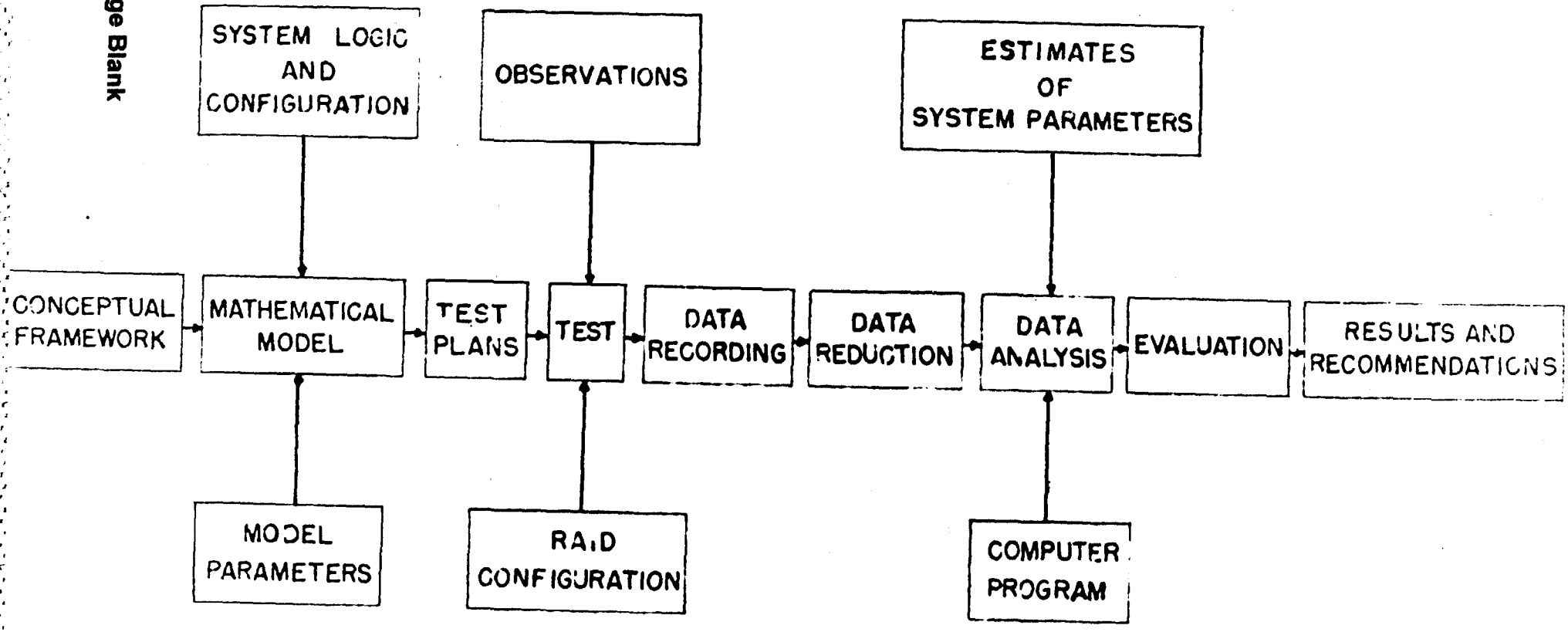


FIG. 2: FLOW CHART OF STEPS INVOLVED IN SYSTEM MODELING

APPLICATIONS OF SELECTING SAMPLE SIZES FOR F-TESTS

Lt. E. L. Bombara
Redstone Arsenal

Often times in rocketry it is necessary to conduct environmental tests on newly developed rounds. For example, it is important to investigate the effects of high humidity on the time required for a certain type of rocket to travel 1000 yards. In addition to determining whether or not high humidity affects the mean time to 1000 yards, it is very important to know how the variance of this time is affected.

In the early stages of testing the desired experiment is a very simple one. A certain number of control rounds (exposed to a standard humidity) and a number of treated rounds of the same type (exposed to high humidity) are to be fired.

The foremost problem confronting the engineer is that of determining how many rounds he should fire in order to obtain reliable answers. He desires a test that will closely predict behavior of the entire population of rockets, but at the same time he can afford to test only the absolute minimum number needed to obtain the required precision of results.

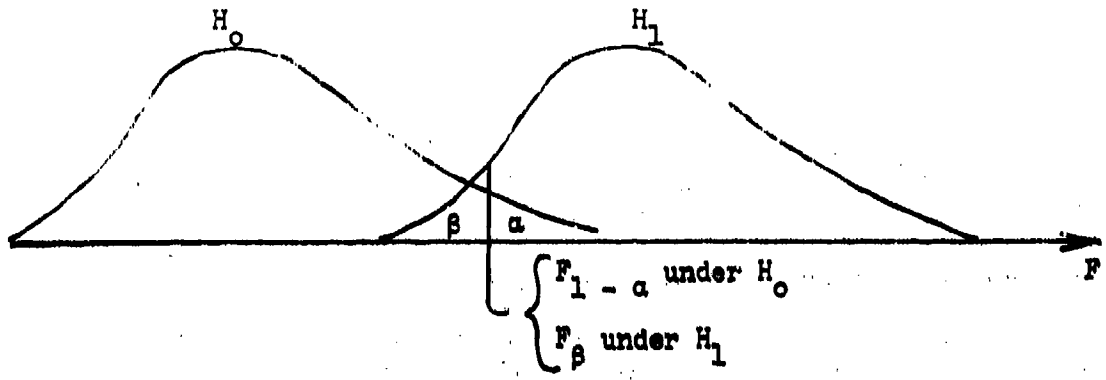
It will be assumed that samples large enough to compare two variances are adequate for comparing the two means.

A well-known method of determining sample sizes for comparing two variances with specified α, β, and ratio of σ_1^2 to σ_2^2 has been developed. Let σ_1^2 be the true variance of the treated rounds, and let σ_2^2 be the true variance of the control rounds. Based on requirements of the parameter in question, it is possible to select a value that is actually not acceptable for the ratio of σ_1^2 to σ_2^2 such that the probability of accepting H_o: $\sigma_1^2 \leq \sigma_2^2$ is β. Let us define this value as λ^2. Now if $\sigma_1^2 \leq \sigma_2^2$, the engineer wants to accept the treated rounds. If $\sigma_1^2 > \sigma_2^2$, he wants to reject the treated rounds and redesign. Given α and β, the theory for obtaining sample sizes is as follows:

The null and alternate hypotheses are

$H_o: \sigma_1^2 \leq \sigma_2^2$

$H_1: \sigma_1^2 > \sigma_2^2$



Let s_1^2 and s_2^2 be unbiased estimates of σ_1^2 and σ_2^2, respectively. Under H_o we have

$$P\left(\frac{s_1^2}{s_2^2} > F_{1-\alpha} \,\middle|\, \frac{\sigma_1^2}{\sigma_2^2} = 1\right) = \alpha$$

and under H_1 we have

$$P\left(\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2} > F_\beta \,\middle|\, \sigma_1^2/\sigma_2^2 = \lambda^2\right) = 1 - \beta$$

where $\lambda^2 > 1$.

Now under H_o

$$\frac{s_1^2}{s_2^2} \sim F_{(\nu_1, \nu_2)}$$

and under H_1

$$\frac{s_1^2}{s_2^2} \sim \frac{\sigma_1^2}{\sigma_2^2} F_{(\nu_1, \nu_2)}$$

$$\therefore \quad F_{1-\alpha, (\nu_1, \nu_2)} = \frac{\sigma_1^2}{\sigma_2^2} F_{\beta, (\nu_1, \nu_2)}$$

and

$$\lambda^2 = \frac{\sigma_1^2}{\sigma_2^2} = F_{1-\alpha, (\nu_1, \nu_2)} \; F_{1-\beta, (\nu_2, \nu_1)}$$

where s_1^2 has ν_1 degrees of freedom, and s_2^2 has ν_2 degrees of freedom. Using an F-Table, ν_1 and ν_2 and, hence, n_1 and n_2 are found by trial and error. Curves have been obtained for $n_2 = k\, n_1$ by plotting sample size against λ^2 for several values of α and β and several values of k. Where α and β are not the same, the curves have been constructed with β less than α. This was done because it is very frequently true in rocketry that the error of accepting bad rounds (those having high variance) is more costly than the error or rejecting good rounds. Notice that $n_1 = n_2$ will produce a smaller total sample size than any other combination.

Two different applications of choosing sample sizes are

1. Suppose σ_1^2 is chosen four times greater than σ_2^2. The engineer is willing to let α be as high as 0.20 but desires β to be no higher than 0.05. Also, he would like to fire three times as many control rounds as treated rounds. To do this, he should refer to the curve for $n_2 = 3n_1$, $\alpha = 0.20$, $\beta = 0.05$ (n_1 will always be the number of treated rounds in this type of problem). Then for $\sigma_1^2/\sigma_2^2 = \lambda^2 = 4$, the curve gives $n_1 = 12$, and $n_2 = 36$. Curves are not yet available for $k = 1/2$, 2, $1/4$, and 4, but linear interpolation between two curves will suffice for these cases.

2. In setting up an environmental test, it is desired to compare means by testing for a source of variation, σ_b^2, among the batches of rockets making up the control rounds. Assuming that a batch consists of m rounds, it is necessary to find the number, b, of batches. The analysis of variance will be of the form

Source	d - f	MS	E(MS)
Between batches	b - 1	s_B^2	$\sigma_e^2 + m\sigma_b^2$
Within batches	b(m - 1)	s_e^2	σ_e^2
Total	mb - 1		

The hypotheses are

$$H_o : \sigma_b^2 = 0$$

$$H_1 : \sigma_b^2 = \gamma\sigma_e^2,\ \gamma > 0$$

Under H_o

$$P\left(\frac{s_B^2}{s_e^2} > F_{1-\alpha} \,\middle|\, \sigma_b^2 = 0\right) = \alpha$$

Under H_1

$$P\left(\frac{\sigma_e^2}{\sigma_e^2 + m\sigma_b^2}\,\frac{s_B^2/s_e^2}{} > F_\beta \,\middle|\, \sigma_b^2 = \gamma\sigma_e^2\right) = 1 - \beta$$

γ is to be chosen in a manner similar to that by which λ^2 was chosen in the preceding illustration. Now

$$\lambda^2 = \frac{E(s_B^2)}{E(s_e^2)} = 1 + m\gamma$$

An approximation for k in the relation $n_2 = kn_1$ is given by

$$k = \frac{b(m-1)+1}{b} \simeq m - 1$$

where $b = n_1$. Having found λ^2 and k, and having chosen α and β, $b = n_1$ can be found from the curves at the end of this paper.

To give a specific illustration of this type of problem, suppose the engineer chooses $\sigma_b^2 = 2\sigma_e^2$. That is, $\gamma = 2$. If, for example, he chooses $m = 4$ rockets per batch (this value is arbitrary), he obtains $\lambda^2 = 1 + (4)(2) = 9$, and $k \simeq 3$. Also, suppose he chooses $\alpha = 0.20$, $\beta = 0.01$. Then referring to the curve for $n_2 = 3n_1$, $\alpha = 0.20$, $\beta = 0.01$, he finds that for $\lambda^2 = 9$, $b = n_1 = 8$.

The reverse procedure of finding m for a particular value of b can be accomplished by trial and error. That is, given specified values of α, β, and γ, different values of m can be selected until the desired value of b is obtained. The curves given here are useful up to values of $m = 6$. Tables 8.3 and 8.4 in Reference 1 may be used for larger values of m and smaller values of α.

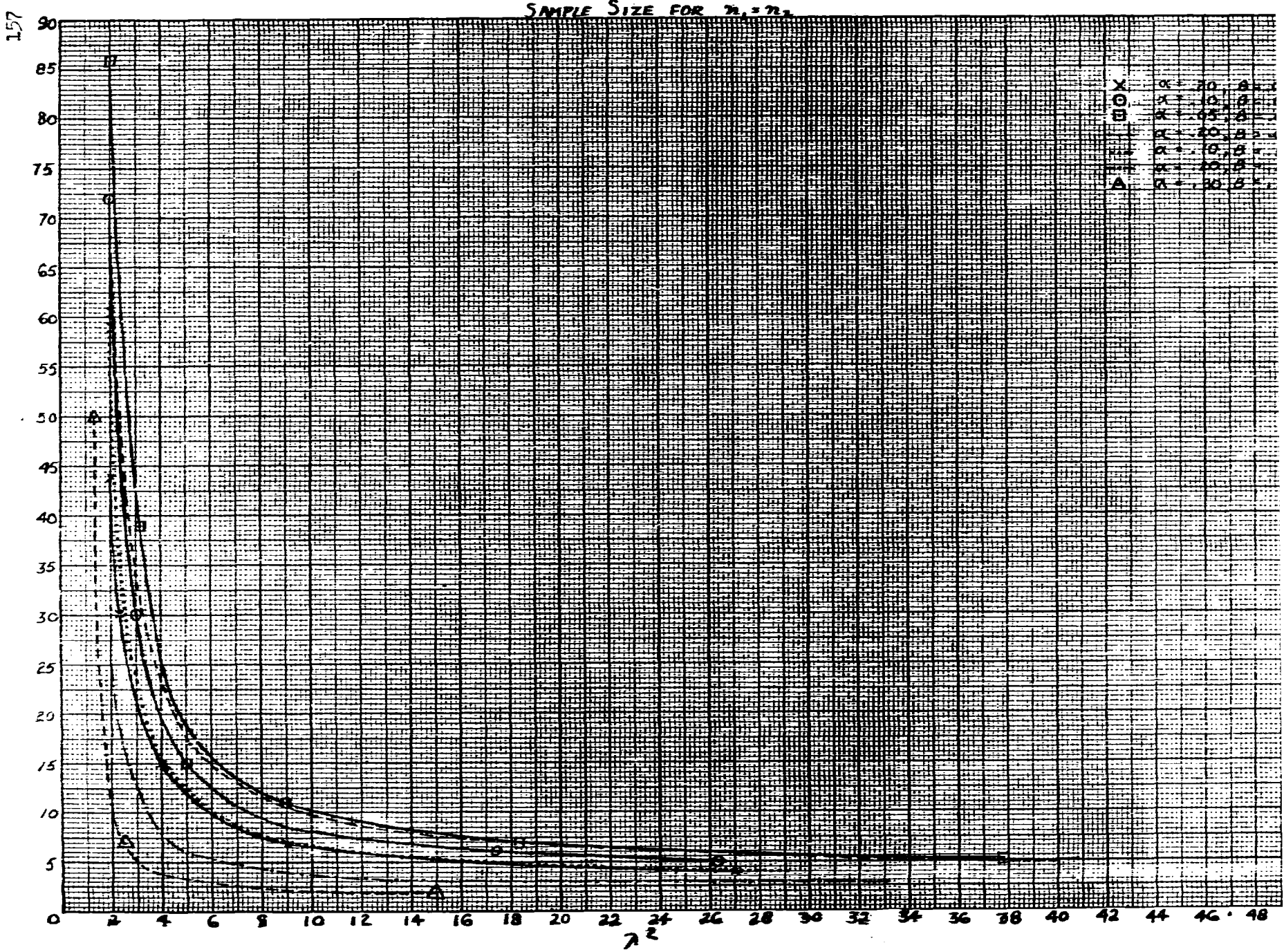
For additional discussion of these two types of problems, along with operating characteristic curves of the F-test, see Reference 2.

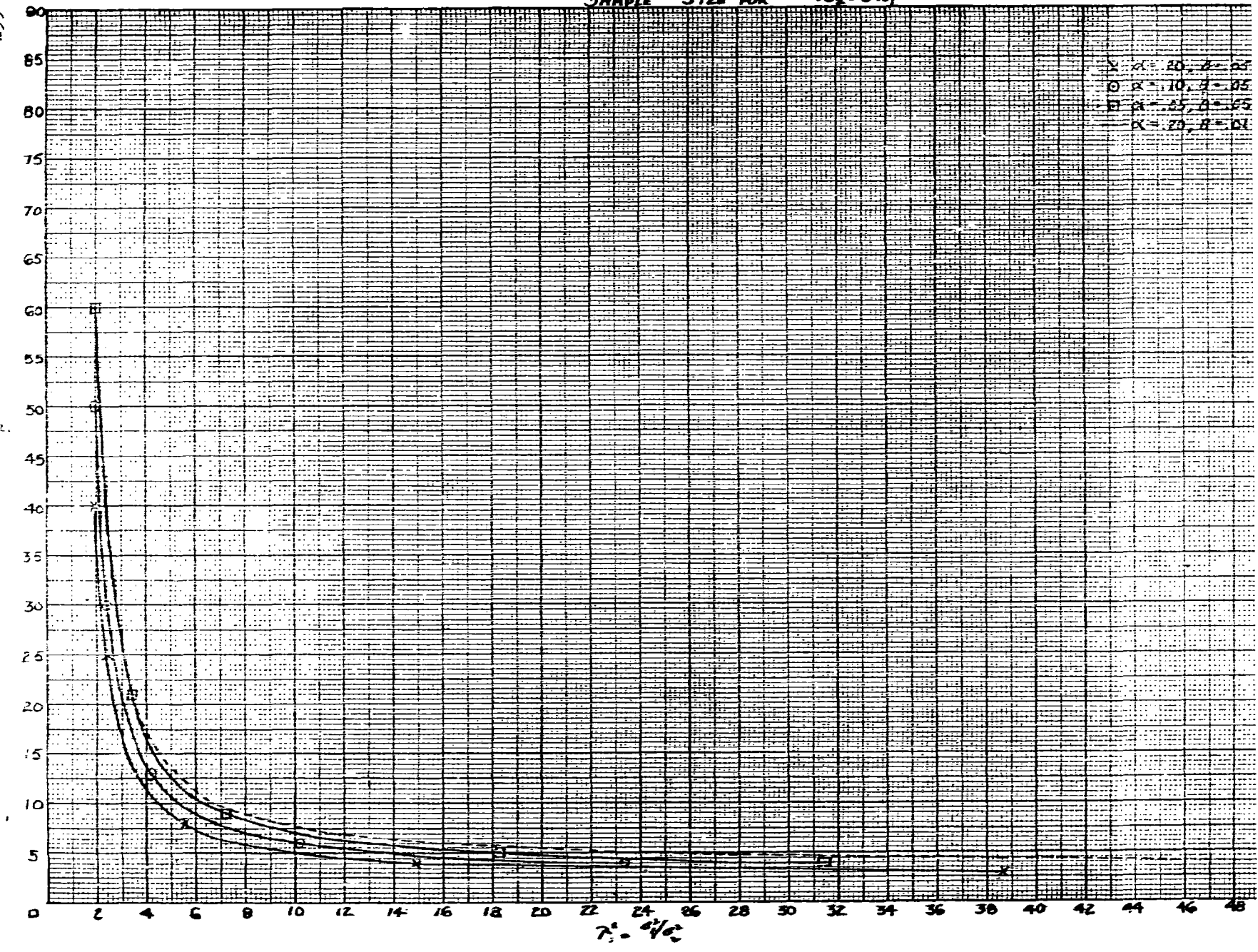
REFERENCES

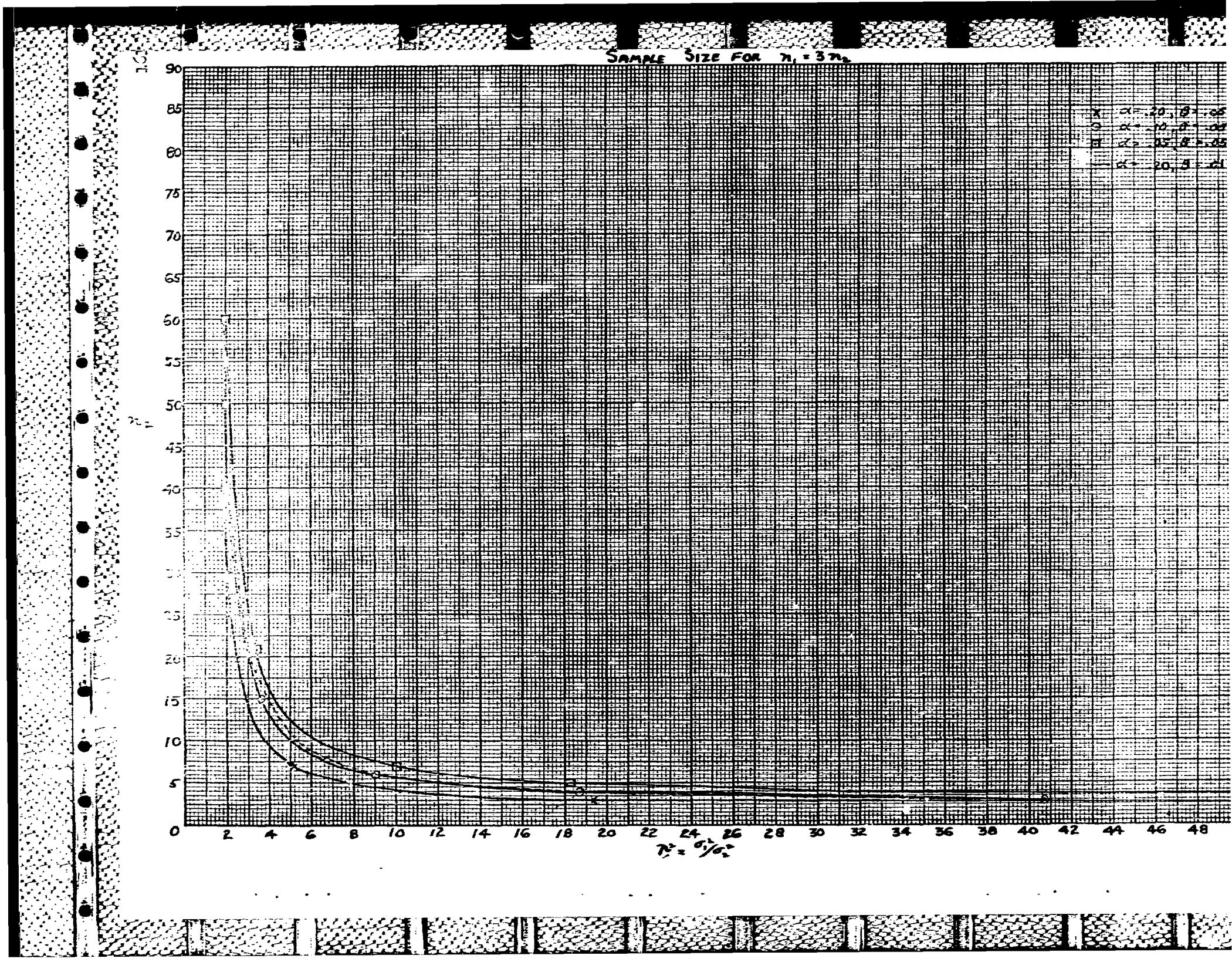
1. Eisenhart, Hastay, and Wallace, Techniques of Statistical Analysis, New York: McGraw-Hill Book Co., Inc., 1947.

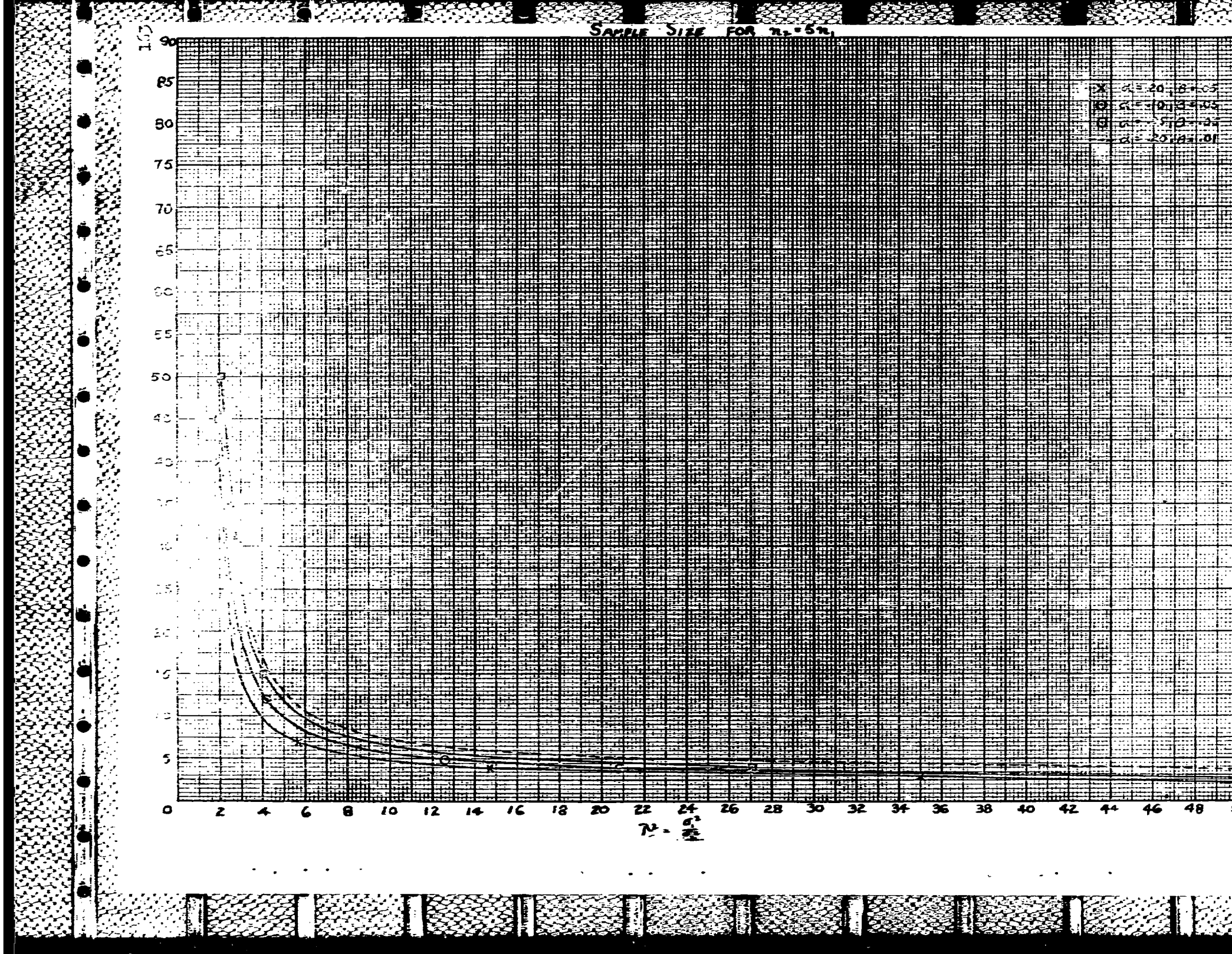
2. Ferris, Grubbs, and Weaver, "Operating Characteristics for the Common Statistical Tests of Significance," Annals of Mathematical Statistics, Vol. XVII (1946), pp. 178-197.

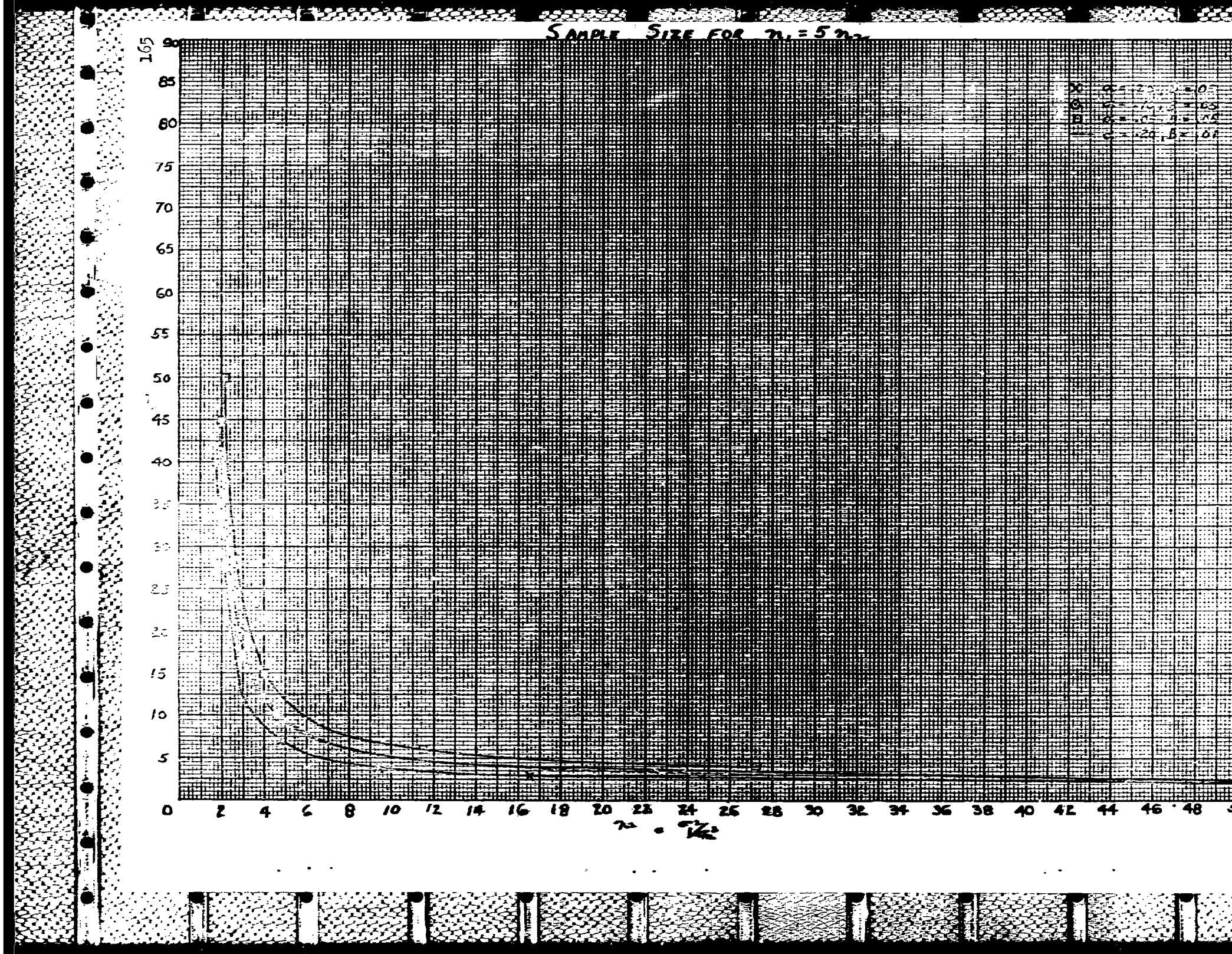
3. A. Hald, Statistical Theory with Engineering Applications, New York: John Wiley and Sons, Inc., 1952, p. 379.


SAMPLE SIZE FOR $n_1 = n_2$
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
λ^2
X α = .20, β =
O α = .10, β =
□ α = .05, β =
α = .20, β =
α = .10, β =
α = .20, β =
△ α = .30, β =


SAMPLE SIZE FOR $n_2 = 3n_1$
n_1
n_2
$\lambda^2 = d^2/\sigma^2$
× α = .20, β = .20
⊙ α = .10, β = .05
⊞ α = .05, β = .05
— α = .20, β = .01
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48


SAMPLE SIZE FOR $n_1 = 3n_2$
$\alpha = .20, \beta = .05$
$\alpha = .10, \beta = .05$
$\alpha = .05, \beta = .05$
$\alpha = .20, \beta = .01$
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
$n_2 = \sigma_1^2/\sigma_2^2$


SAMPLE SIZE FOR $n_2 = 5n_1$
n_1
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48
$N = \frac{\sigma_1^2}{\sigma_2^2}$
X $\alpha = .20, \beta = .05$
O $\alpha = .10, \beta = .05$
□ $\alpha = .05, \beta = .05$
— $\alpha = .20, \beta = .01$


SAMPLE SIZE FOR $n_1 = 5n_2$
165
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
$n_2 = \frac{n_1}{5}$

RECOMMENDATIONS FOR THE DESIGN OF EXPERIMENTS FOR ESTIMATING QUADRATIC REGRESSION

Pvt. Paul G. Sanders
Redstone Arsenal

INTRODUCTION. The following problem is treated. A total of N observations, $y_1, y_2, \cdots, y_N$, may be taken at any locations, x_j in the range x_L to x_H. The y_i are uncorrelated and have common variance V_0. The relation between y and x is

$$E(y_i) = a + bx_i + cx_i^2$$

where E ( ) stands for the expectation or mean value of the symbol in brackets. It is desired to select values of x at which to observe y so that certain specific questions about the relation, Eq. 1, may be answered as efficiently as possible. Best spacings of the x are given for the following situations:

1. Interpolation, to minimize the maximum standared error of the estimate $\hat{y}(x)$ in the range x_L to x_H.

2. Extrapolation, to minimize the standard error of $\hat{y}(x_0)$ for some $x_0 > x_H$.

3. Testing H_0: $c = 0$.

The situation described is one frequently encountered by experimenters in engineering or laboratory work. The variable x often represents pressure or temperature, with limits x_L and x_H prescribed by equipment restrictions.

The estimation of the constants, a, b, and c, is made by least squares methods which provide expressions for the standard errors, each a function of the x's selected, that are needed in answering situations 1, 2, and 3. The least squares estimates are denoted by the symbol $\hat{}$.

The recommended spacings (set of x's for an experiment) depend upon a result of Garza[1] which implies, for our problem, that exactly three distinct values of x will suffice for any problem like those above. Thus, all of the best spacings consist of three values, x_1, x_2, and x_3, satisfying $x_L \leq x_1 < x_2 < x_3 \leq x_H$ with n_1, n_2, and n_3 ($\sum_{j=1}^{3} n_j = N$) observations of y at the corresponding values of x. Values of n_j will generally not be integers; care must be taken in rounding calculated spacings off to integer values. For small N, a fine-structure study may be required.

This result simplifies both the problem of selecting a spacing and the actual calculation of the constants in a given experiment. Denote by $\bar{y}_j$ the arithmetic mean of the n_j measurements of y and x_j. Then it can be shown that the least squares estimate of Eq. 1 passes through the three points $(x_j, \bar{y}_j)$, $j = 1, 2, 3$. Thus the least squares estimate can be written in the Lagrange form

$$\hat{y}(x) = \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)} \bar{y}_1 + \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)} \bar{y}_2 + \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)} \bar{y}_3 \tag{2}$$

which may be written

$$\hat{y}(x) = \sum_{j=1}^{3} F_j(x)\bar{y}_j \tag{3}$$

with an obvious notation. We have at once

$$\text{Var}\left[\hat{y}(x)\right] = V_0 \sum_{j=1}^{3} \left[F_j(x)\right]^2 \frac{1}{n_j} \tag{4}$$

We now proceed to consider the three problems separately.

INTERPOLATION. In this case, we wish to know Eq. 1 as well as possible over the range x_L to x_H. This may be done by finding a spacing which minimizes the maximum $\text{Var}\left[\hat{y}(x)\right] = V_{max}$ for $x_L \leq x \leq x_H$. A simple derivation of this spacing was given by Garza[1]. Note from Eq. 4 that

$$\text{Var}\left[\hat{y}(x_j)\right] = \frac{V_0}{n_j} \leq V_{max}$$

Then

$$\frac{V_0}{3} \sum_{j=1}^{3} \frac{1}{n_j} \leq V_{max} \tag{5}$$

The minimum value of $\sum_{j=1}^{3} 1/n_j$ constrained by $\sum n_j = N$ is $9/N$ with $n_j = N/3$.

Hence,

$$\frac{3V_0}{N} \leq \min V_{max} \tag{6}$$

Now if the symmetric spacing $x_1 = x_L$, $x_2 = (x_L + x_H)/2$, $x_3 = x_H$, and $n_1 = n_2 = n_3 = N/3$ is used, Eq. 4 has a differentiablemaximum at x_2 equal to $3V_0/N$. At both x_L and x_H, Eq. 4 is increasing as the end of the interval is approached from within, and Eq. 4 is equal to $3V_0/N$ at the ends. Therefore, $V_{max} = 3V_0/N$. Hence, the spacing is the desired one since it produces the equality of Eq. 6. Thus the best spacing for interpolation is to take N/3 observations at each of x_L, $(x_L + x_H)/2$, and x_H.

The problem is solved, but a few comments are in order. An important objection to the "best" spacing is that it offers no information about inadequacies in the model, Eq. 1. Thus we wish to examine the sensitivity of the best spacing to variations which allow tests of the adequacy of the quadratic model. Consider the spacings S_1 and S_2, where $x_L = -1$, $x_H = 1$ are used for simplicity and $p_j = n_j/N$.

S_1	p	1/4		1/8		1/4		1/8	1/4
	x	-1		-1/2		0		1/2	1
S_2	p	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
	x	-1	-5/7	-3/7	-1/7	1/7	3/7	5/7	1

The ratio of the standard errors of the two spacings to the standard error of the best spacing is shown for several values of x. (Standard error = $\sqrt{\text{Var}\left[\hat{y}(x)\right]}$)

x	0	$\pm$0.2	$\pm$0.4	$\pm$0.6	$\pm$0.8	$\pm$1.0
S_1	0.93	0.93	0.95	0.97	1.05	1.11
S_2	0.87	0.87	0.89	0.97	1.20	1.37

It appears that S_1 compares quite well with the best spacing, but S_2 is exceedingly weak at the ends. This indicates that for large N the often-used practice of taking observations an equal distance apart sacrifices much accuracy at the end points.

In conclusion, the recommended procedure is to use a spacing which will allow detection of inadequacies in the quadratic model but which does not greatly increase V_{max}.

EXTRAPOLATION. If the experiment is performed to determine the estimate $\hat{y}(x_0)$ for x_0 outside the interval (x_L, x_H), the problem is one of extrapolation. We shall assume that $x_0 > x_H$, but the solution for $x_0 < x_L$ can be obtained easily from the solution for $x_0 > x_H$.

A brief justification is given for the results. Equation 2 becomes

$$\hat{y}(x_0) = \sum_{j=1}^{3} f_j \bar{y}_j \tag{7}$$

where

$$f_j = F_j(x_0)$$

Then

$$\text{Var}\left[\hat{y}(x_0)\right] = V_0 \sum_{j=1}^{3} f_j^2 \frac{1}{n_j} \tag{8}$$

Then n_j, which minimize Eq. 8, are

$$\frac{n_j}{N} = \frac{|f_j|}{\sum_{j=1}^{3} |f_j|} \tag{9}$$

The values of x can be shown to be, again

$$x_1 = x_L,\ x_2 = \frac{x_L + x_H}{2},\ x_3 = x_H \tag{10}$$

Thus, to minimize Eq. 7, n_j is determined from Eq. 9 with x_j given by 10.

As an example, suppose $x_1 = x_L = 10$, $x_3 = x_H = 20$, $x_2 = 15$, and $x_0 = 25$. Then

$$n_1 = \frac{N\dfrac{(25-15)(25-20)}{(10-15)(10-20)}}{\dfrac{(25-15)(25-20)}{(10-15)(10-20)} - \dfrac{(25-10)(25-20)}{(15-10)(15-20)} + \dfrac{(25-10)(25-15)}{(20-10)(20-15)}}$$

$$= 0.143N$$

$$n_2 = 0.428N$$

$$n_3 = 0.429N$$

Then, from Eq. 8

$$\text{Var}\left[\hat{y}(x_0)\right] = \frac{49V_0}{N}$$

If we want this equal to, say the variance of a single observation, V_0, then N = 49 observations will be required. This number seems quite large when it is remembered that the same precision can be attained from x_L to x_H with just three observations.

In summary, in planning an experiment where extrapolation is unavoidable, even the best spacing often requires a large number of observations; simple calculations like these should be made to determine beforehand what may be expected from proposed extrapolation.

TESTING THE HYPOTHESIS ABOUT c. For large x_0 Var $[\hat{y}(x_0)]$ approaches its dominant term, Var $(\hat{c})x_0^4$. Hence, for large x_0, when Var$[\hat{y}(x_0)]$ is minimized, Var (c) is also minimized. Letting x grow large in Eq. 8. we find

$$n_1 \to \frac{N}{4},\ n_2 \to \frac{N}{2},\ n_3 \to \frac{N}{4} \tag{11}$$

This is the spacing that minimizes Var $(\hat{c})$; hence it yields the best test for hypotheses about c. The minimum value of Var $(\hat{c})$ is

$$\text{Var}\ (\hat{c}) = \frac{64V_0}{N(x_H - x_L)^4} \tag{12}$$

SUMMARY. Best spacings have been given for three common situations in quadratic regression. The spacings are often different from those commonly used. They depend on obtaining several independent observations at the same value of x. Where the cost of an observation is independent of x, these spacings are minimum cost spacings. More general considerations in the design of regression experiments are found in References 1, 2, and 3. Reference 4 gives best spacings for estimating straight lines. Reference 5 gives some of these results together with a discussion of rate of subsampling.

BIBLIOGRAPHY

1. de la Garza, A., "Spacings of Information in Polynomial Regression," Annals of Math. Stat., 25 (1954) pp. 123-130.
2. Elfving, G., "Optimum Allocation in Linear Regression Theory," Annals of Math. Stat., 23 (1952) pp. 255-262.
3. Chernoff, H., "Locally Optimum Designs for Estimating Parameters," Annals of Math. Stat., 24 (1953) pp. 586-602.
4. Daniel, C., and Heerema, N., "Design of Experiments for Most Precise Slope Estimation of Linear Extrapolation," Journal of the American Statistical Association, 45 (1950) pp. 546-556.
5. de la Garza, A., Hawyhurst, D., and Newman, L. T., "Some Minimum Cost Procedures in Quadratic Regression," Journal of the American Statistical Association, 50 (1955), pp. 178-184.

A WIDE BAND TELEMETERING SYSTEM

R. A. Parkhurst
Diamond Ordnance Fuze Laboratories

In making chaff reflection studies several methods have been employed. As examples, one method synthesizes chaff by using randomly spaced pins in waveguide. Another involves dropping chaff piece by piece in still air, making reflection measurements and integrating all such data into the composite signal which would occur if all pieces were dropped simultaneously.

Reflection studies are not only difficult due to the problem of synthesizing chaff echo signals but are also further complicated by the type of signal being reflected. If a cloud of chaff is in the air and a Pulse radar beam is swept through it, the echo amplitude and stretching will be one amount when the beam is aimed directly at the chaff, but when the beam strikes only the side of the chaff cloud, the stretching and amplitude will be of a different value.

In the event that cw is used instead of pulses the reflection will vary with respect to the position of the chaff in the antenna pattern, or the rate at which the chaff is passing through the pattern.

To synthesize such conditions in the laboratory is quite difficult, if not impossible, and mathematical analyses become so complex that they produce little more than very general results.

Two methods are available for obtaining genuine reflections from chaff. One of these is to use a supersonic sled facility, several of which are available at test stations throughout the country. In this type of test a fuze is mounted on a sled which is driven by rocket motors and reaches speeds up to 2000 feet per second. Various targets may be placed along the side of the track and signals from them will be telemetered to the receiving station. For chaff studies, chaff may be dispensed over the track, either by aircraft or by mortar shell.

Several complications are involved in this type of test. The main one is that the fuze must be made insensitive to ground echos. This requires altering the fuze radiation pattern on the side towards the ground which may cause distortion of the signal either by general mishaping of the pattern or by producing erroneous signals in the fuze due to the imperfection of the shielding material used. Multiple reflections generated between the chaff and the ground may also be present, and these can also create errors in the data obtained.

The most realistic method of obtaining chaff reflections is to fire a test vehicle past chaff in the sky. Tests of this nature have been performed and good results have been obtained. These tests were set up for a specific purpose; namely, chaff reflections as seen by one type of fuze, and so the data obtained generally is applicable to only one type of fuzing. The method used in setting up and performing these tests is quite similar to a regular missile flight test except that special fuze telemetering is employed, and a drone with a chaff dispenser is used.

Figure f is a diagram of a typical arrangement for a chaff test. The shaded area is the ocean firing range. The control stations, telemetering ground stations, and landing fields are located on the land at the right. The target plane flies the marked course. It's air speed is 310 feet per second and it carries what may be called a standard chaff dispenser.

The launch aircraft follows a similar course and thus makes a tail approach to the target. Its velocity is up to 660 ft./sec, and the timing check points are to assure proper location of the planes during the test. The plane locations are plotted on radar plotting boards at the control station, and if either the drone or launcher is not at the prescribed point at the proper time, corrective directions are given to bring them back on course. Both planes fly at 8000 feet altitude, and the missile is launched at the point marked 0 time, about 3,500 yards from the target.

A camera plane flies 2,000 feet to the left of the launch aircraft and slightly down and to the rear. This plane carries cameras which are boresighted with the planes guns, thus being aimed by the gunsights. Other cameras take pictures through the windshield, and in some cases hand held cameras are used for extra coverage.

The launch aircraft also has several cameras. One has a telephoto lens for close up pictures of the intercept. One is boresighted to the launcher on the plane and sees the missile and target with respect to that angle. A third covers the operation through the windshield.

When a test is performed, the drone, launcher, camera plane, and other test aircraft take off at -30 minutes. A TM check is made with the ground station, and a dry run is made against the target. The planes are then repositioned and the test proceeds. Positions and direction are called out by the radar control station. At -1 minute a final position check is made and if all is well, all operation personnel are notified. At -6 seconds all TM equipment is started in the ground station. At -3 seconds the photo plane cameras are started and the chaff dispenser in the drone is turned on. At -1.5 seconds the launch plane cameras are started and at 0 time the pilot fires the missile. At +15 seconds the photo plane cameras stop and at +25 seconds the launcher cameras stop. A post launch check is made with the launch plane making runs against another drone with a pilot in it for radar calibration purposes. Cameras are also placed in wing pods on the drones to obtain more precise intercept data.

The physical execution of the test is only part of the entire operation. Success or failure depends on successful operation of aircraft and missile, skillful performance of operating personnel, and proper operation of a rather complex telemetering and recording system.

In order to obtain signals from clouds of chaff, a test vehicle with a fuze must be launched and guided through the chaff dispensed by the target aircraft. As the missile passes by the chaff and target, any signal delivered by the fuze receiver is fed to the telemeterine system. Thus, on one test, echos from several clouds of chaff and one target are obtained.

---

* Figures appear at end of the article

A special telemetering system is used which transmits the signals to the ground station. Calculations have shown that frequency components from dc to 100 kc may be present and that the TM system should have a bandwidth as wide as this. Not only must this relatively large bandwidth be accommodated, but the airborne portion of the system must be able to wi. stand violent vibration. The transmitter, when subjected to vibration equivalent to that expected in flight, must not fail mechanically, and it must not generate spurious noise which could be confused with the desired signals.

In order to meet these telemetering requirements two approaches were possible. One, to develop an entire system which fully met the requirements, was rejected as being too time consuming and expensive. The other was to select and use any commercially available equipment which most nearly met the needs.

Figure 2 shows a standard multichannel FM-FM telemetering system which was investigated for usable components and techniques. This system consists of a crystal controlled VHF transmitter which is either phase-modulated or frequency-modulated by subcarriers. The transmitter accepts modulation up to 100 kc, and the subcarriers have various frequencies ranging from a few hundred cycles up to 70 kc. Each subcarrier is frequency modulated with a signal which is to be telemetered. The subcarrier frequencies are chosen so that then each is deviated ± 15% from its center frequency, none of the modulation sidebands will interfere with the subcarriers adjacent to it in the spectrum. Also, any harmonics generated by non sinusoidal subcarrier oscillators are filtered out before applying the signal to the transmitter in order to avoid interference.

The ground station has a VHF receiver which is tuned to the transmitter frequency. The output of this receiver is fed into a bank of filters and discriminators so that each subcarrier is separated out and fed to a discriminator tuned to its own center frequency. The discriminator demodulates the subcarrier, and in this manner the information applied to each subcarrier is recreated in the ground station.

The center frequency of each subcarrier more or less determines the bandwidth of the signal which may be applied to it. (For instance, if a 40 kc subcarrier is deviated ± 15%, the modulation frequency allowable for an index of 5 would be 1,200 cycles. Deviations of greater than ± 15% would create side bands which would interfere with adjacent subcarrier signals, and to maintain low signal to noise ratios it is advisable to keep modulation indices above 5. Thus, in normal usage the maximum frequency applied to a 40 kc subcarrier is 1,200 cycles.) The greatest bandwidth available for a subcarrier is dc to 20 kc. This may be obtained with a 70 kc subcarrier and using a modulation index of 1. This not only lowers the signal to noise ratio considerably, but requires modification of the subcarrier and the subcarrier discriminator.

This maximum of 20 kc was not sufficient for the chaff tests, so it was decided to investigate direct modulation of a transmitter.

A phase modulated crystal controlled transmitter was checked for suitability. In a transmitter tested, a crystal oscillator ran at about 20 mc, and this frequency was multiplied up to the desired carrier frequency by several multiplying stages. One of the multiplier tuned circuits was tuned by a reactance tube, the reactance of which was varied by the modulating signal. This, in turn produced a phase lead or lag in that particular stage and thus phase-modulated the carrier at that point. The phase modulation was then multiplied along with the carrier until the output at the desired frequency was obtained with its phase varying proportionally with the modulation signal.

In this type of transmitter, if a linearly rising signal is applied as modulation, the phase will advance continually at a linear rate. If the carrier is observed during this period, it will be noted that as long as the phase advances, the frequency will be increased. That is, the carrier will be some steady value above its normal unmodulated frequency. If this signal is detected in an FM discriminator or ratio detector, we will get a constant voltage output. The modulation applied, however, is a sawtooth, or linearly rising signal, so it is apparent that a phase-modulated signal, when detected by an FM receiver, will be differentiated.

A further example of this would be to apply a square wave to the transmitter. When the input changes from its negative value to its positive, the phase of the carrier is shifted by some amount depending on the amplitude of the applied signal. As long as the input voltage remains constant, as during 1/2 cycle of the square wave, the carrier remains at its unmodulated frequency, but advanced or retarded in phase. In this case, a discriminator or ratio detector would see the same frequency at all times except when the carrier was shifting from one phase reference to another. During these shifts, the discriminator would deliver pulses proportional to the rate of change in phase. These pulses are, again, the derivitive of the applied modulation signal. Figure 3 shows clearly this differentiating action between the input and output of a PM transmitter with square wave modulation applied. In normal FM-FM usage this differentiating action is of little importance since the information being conveyed is strictly the individual subcarrier frequencies and not their waveforms. For wideband purposes, namely in the desire to preserve wave forms, this characteristic is quite a hinderance.

Figure 4 shows another feature of PM transmitters; namely the sloping frequency response curve. Since the frequency generated by shifting the carrier is what the receiver detects, the slower the phase is shifted, the lower will be the effective deviation. This, in effect, holds the modulation index constant, which results in the characteristic that as the modulation frequency decreases, the amplitude of the detected signal decreases. This creates a frequency versus amplitude response which is quite poor when compared with that of an FM transmitter. The frequency response below one kc is generally too low to be useful, and applying larger input voltages at the lower frequencies results in severe distortion. This diagram shows flattening of the response curve above 10 kc which is due to an integrating network across the input. This effectively attenuates the high frequencies, thereby reducing the modulation index as the frequency is increased. One

method for increasing low frequency response is frequency-modulating the crystal at low frequencies. The crystal can not be "pulled" very far, but dc response has been obtained and a curve as shown by the dotted line was achieved. Frequency modulated transmitters of the non-crystal-controlled type were also tested. In the past, TM equipment manufacturers have produced many FM transmitters for use in FM-FM TM systems. With more and more tests being conducted simultaneously and more and more data on the air, crystal control to keep one transmitter from drifting into another's channel has become a must, and non-crystal-controlled transmitters have fallen into general disuse.

For wideband use the major drawback in these transmitters, aside from lack of crystal control, was their lack of rigid construction. In almost all cases when the transmitters were subjected to vibration such as to be encountered during the test, noise would be generated in quantities equal to or greater than the signal being telemetered. Several transmitters of this type have been strengthened mechanically and vibration tested. From a small group of mechanically sound transmitters several test records have been obtained which have not been bettered by any other transmitter. Figure 5 is a response curve of an FM transmitter. The frequency response of an FM transmitter is quite good. By modifying the input circuitry, dc response can be obtained in some models.

Figure 6 shows that phase-shifting of the modulating signal is quite low and that good output wave form fidelity is maintained. The square wave response of an FM transmitter is shown in this diagram. This good frequency-response and phase-response also applies to crystal stabilized transmitters, which are more desirable for both mechanical and stability reasons.

In a crystal stabilized FM transmitter, an oscillator on the order of 30 mc is modulated with a reactance tube, and its frequency multiplied up to VHF region. At the point where the oscillator frequency is doubled, a portion of it is mixed with a signal from a crystal oscillator, and the difference frequency is fed to a discriminator. The output of the discriminator is returned to the modulating reactance tube and thus tries to keep the difference between the transmitter oscillator and crystal oscillator at a fixed amount. The output frequency is thereby maintained fixed at almost crystal accuracy. The degree to which the oscillator frequency is held constant depends upon the frequency response of the correction network from the discriminator to the reactance tube. If a very long time constant is employed in this feed back loop, it will take a relatively long time for the discriminator to shift the oscillator back to its center frequency after a dc step has been applied to the modulation terminals. With this type of system it is evident that dc response can never be achieved, but response down to a few cycles is readily attained.

The particular transmitter tested was quite insensitive to vibration. Of all types checked, it had the most desirable characteristics with a minimum of extra work necessary. This type of transmitter was selected

for use in the final TM system. One drawback found later was that in the transmitter selected, the tubes were run over rating and thus had a considerably shortened life. This didn't matter too much during tests since a flight lasts only a few minutes, but during the hours of TM calibration and checking, at least one transmitter has been run beyond its useful life.

Other factors involving the selection of transmitters or, for that matter, any item, is the environment in which it has to live. An example of what can happen is pointed up by a transmitter build ty DOFL and flown in a 5" rocket. It was not at all ruggedly constructed and when flown, was surrounded by one inch thick foam rubber. This transmitter produced almost noise free records from many rocket firings, yet when tested on a vibration table under conditions expected to be encountered in our tests, it was not only extremely noisy, but rapidly fell apart.

The vibration test given to all transmitters was ten g's in three planes from 20 cycles to 500 cycles. The output of a receiver tuned to the transmitter frequency was observed during the shake test and, with no input, the output was to remain at less than 5% of the output observed with a maximum allowable modulation signal. If more than 5% noise was observed, the transmitter was rejected.

After selecting the crystal stabilized FM transmitter, it was necessary to find suitable ground station equipment. The receiver was by far the easiest part to choose in setting up the wideband TM system. A standard VHF telemetering receiver was checked for frequency response and found to be good from a few cycles to 100 kc. In the event that dc response is eventually obtained in a transmitter, the receiver can easily be modified to deliver dc.

After transmitting the signal to the ground and detecting it, the problem was to record it so that it could be regenerated electrically. From photographic film, this is quite impractical, if not presently impossible.

Wideband tape recorders are made for standard FM-FM systems and are available with bandwidths from 200 cycles to 80 or 90 kc. A typical tape recorder response curve is shown in Figure 7. As video recorders, these machines produce a differentiating action similar to a PM transmitter since when a tape is played back, the voltage generated is proportional to the rate of change of flux on the tape. Figure 8 shows a tape recorder phase distortion. There is also a more or less mechanical phase distortion; this is produced by the phenomena that as the recording frequency is increased, the position of the maximum flux in the recording head gap will move in its relative position, locating itself physically closer to one of the poles.

The net result is that although a tape recorder has fairly good frequency response, and can be modified to go down to 50 cycles, it has relatively poor phase fidelity and will distort wave forms with high harmonic content rather severely.

There are available carrier type recording systems which frequency-modulate a carrier and record this carrier on the tape. This system produces a practically flat response from dc to 10 kc. By altering existing units and sacrificing some of the 40 db signal to noise ratio, bandwidths up to 20 kc can be obtained.

A very promising device in the recording field is a new video recorder designed for television use. It has a flat frequency response from 20 cycles up to 4 megacycles. This device, using a one Mc carrier system, for instance, could easily record from dc to 100 kc. At the time of our testing, the video recorder was not available, and the previously mentioned carrier system did not have enough bandwidth, so a standard FM-FM system recorder was used. The signals received turned out to be low in harmonic content, so it was felt that little distortion was present. Also, the lack of dc response was partially compensated for by recording the signal at the receiver directly on film. This gave a visual record of the signal down to the low frequency limit of the transmitter. Film records are also made for visual inspection of signal wave forms. During the test the signal from the receiver is photographed on both high speed and slow speed films.

There are problems encountered in attempting this direct recording. First of all, if fair resolution of the signal is desired, the film speed must be a minimum of 400 inches per second. For good resolution the speed must be greater, approaching 100 feet per second. In this event, to cover a ten-second test, a 1000 ft. camera capacity would be required, and available Fastax or Eastman high speed cameras hold only a hundred feet of film. Perhaps with extremely accurate timing, the precise second of encounter could be recorded, but no reliable method for starting the camera at the right moment is available.

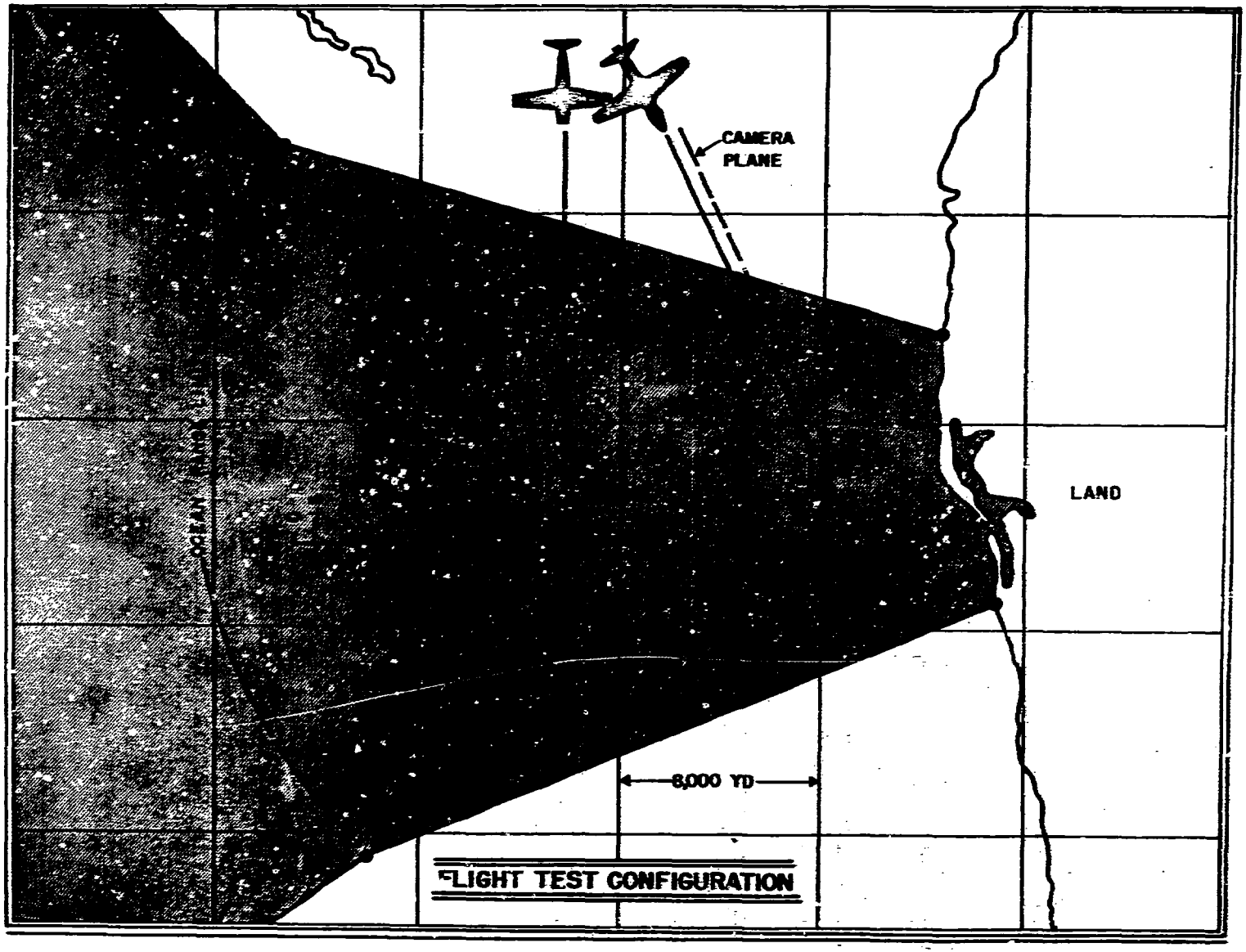
A Miller high speed oscillographic recorder is available which records on photographic paper. This device has a paper speed of 400 inches per second and a capacity of 12 seconds running time. Although the speed is lower than that desirable for good visual records, usable records are practically guaranteed with a minimum of timing problems. This machine is presently being modified to run at 800 inches per second, and the film magazine capacity is being doubled.

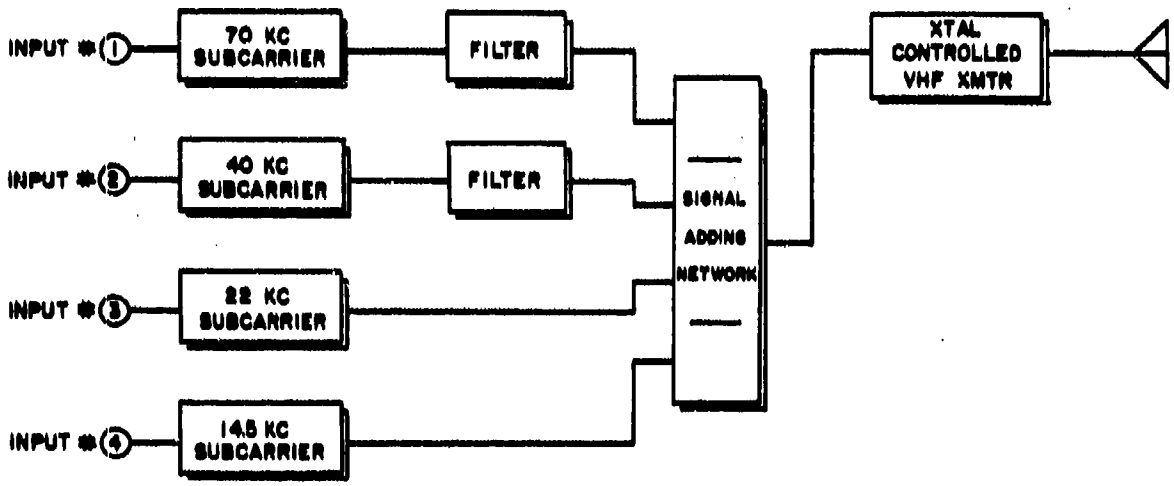
Records are also made on 35 mm. film running at 60 inches per second. These films show overall characteristics and existance of signals, but are not of too much value for anything else since wave forms cannot be distinguished, frequencies cannot be measured, and they can not be "played back" electrically.

Figure 9 shows the entire wide band telemetering system. This consists of a crystal stabilized FM transmitter which is modulated with the signal from the fuze receiver. The particular transmitter chosen was selected for its good frequency response and freedom from microphonics. The receiver is a standard FM receiver. The signal is recorded on tape for playback analyses. It is also recorded on high speed film for visual

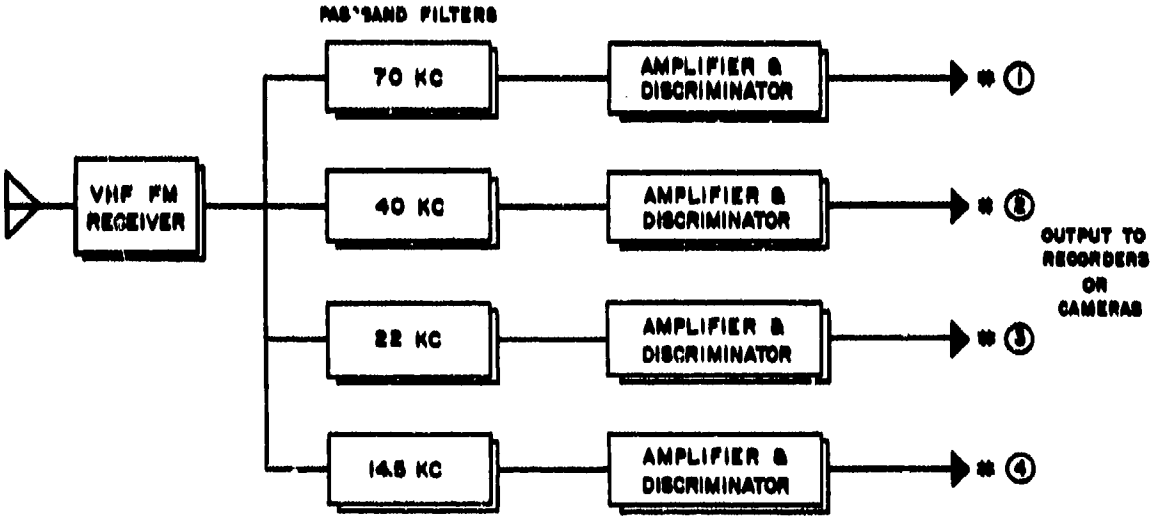
waveform analyses and on slow speed film for visual inspection of general overall envelope structure.

Good results have been obtained from several flight tests in which characteristics of chaff echos have been easily discernible from aircraft echos. Unfortunately, since these differences in characteristics apply to specific fuses, they cannot be discussed in this paper.

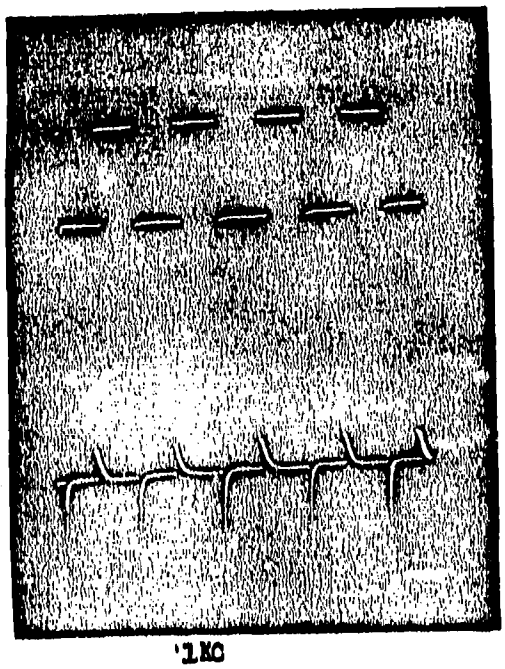

CAMERA
PLANE
LAND
8,000 YD
FLIGHT TEST CONFIGURATION



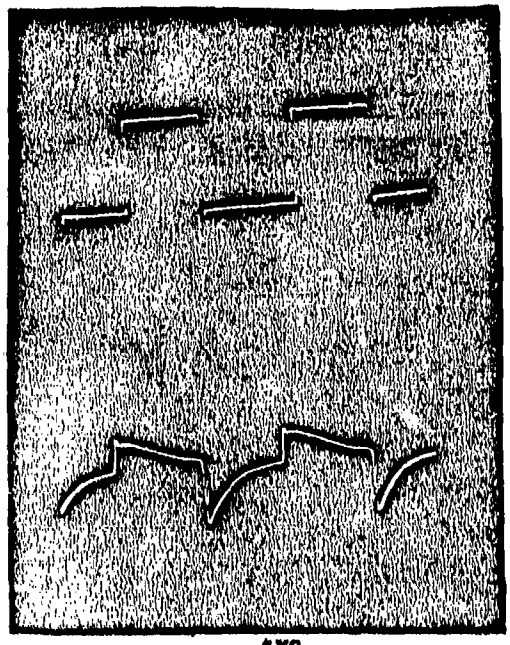
TM TRANSMITTING CONFIGURATION



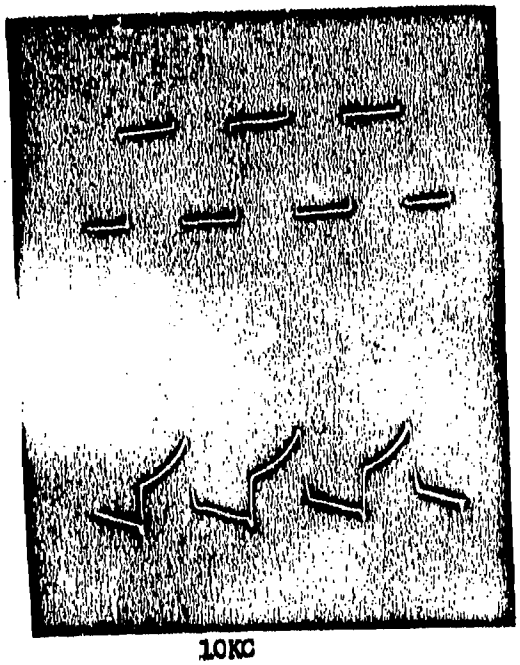
FM-FM TELEMETERING SYSTEM



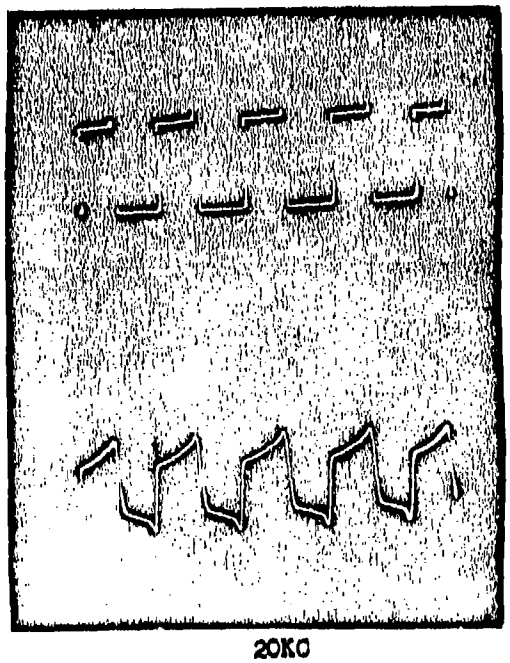
1KC



5KC

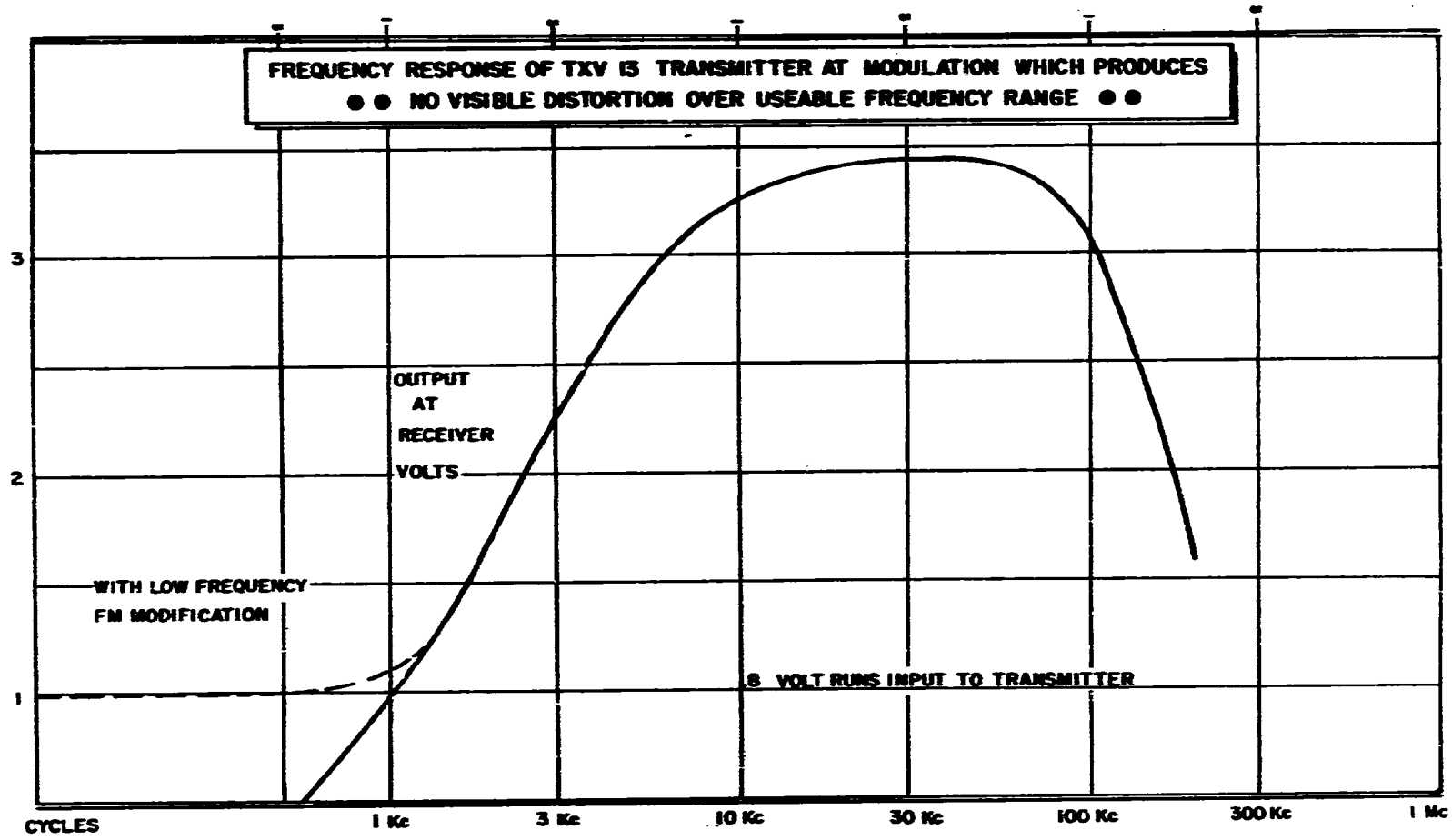


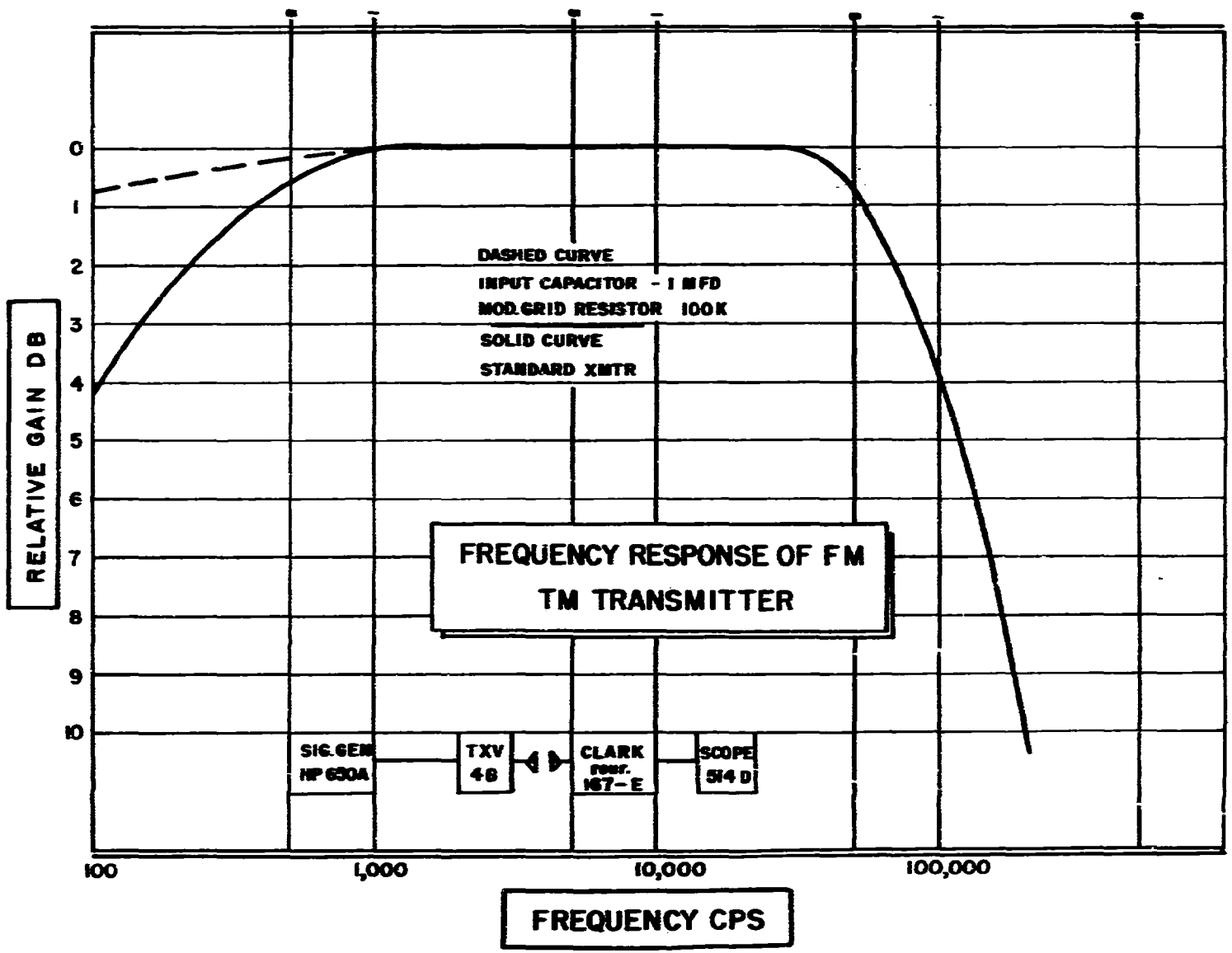
10KC

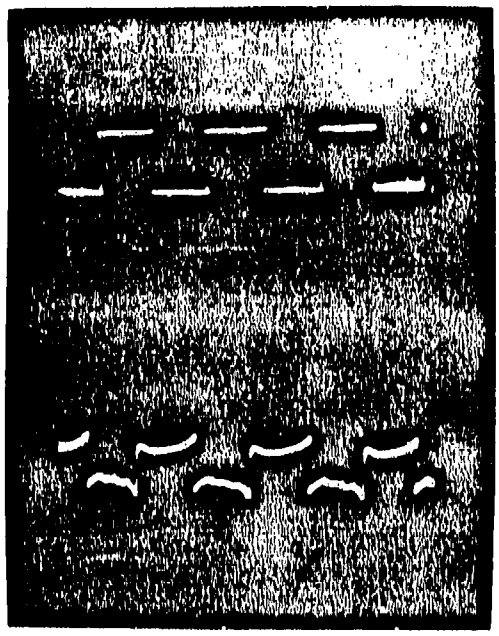


20KC

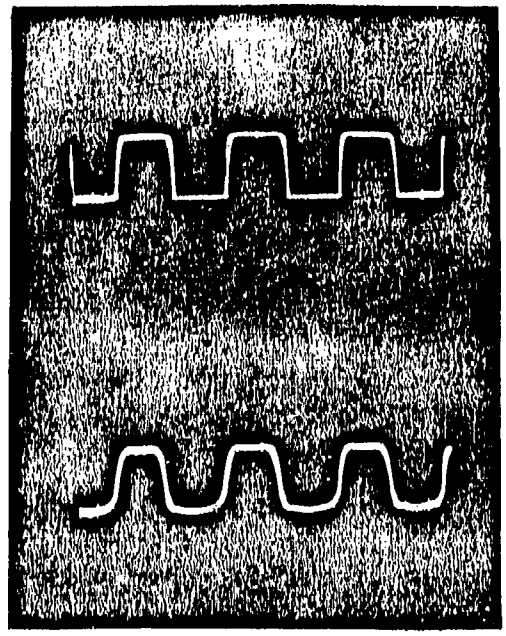
SQUARE WAVE RESPONSE OF PHASE MODULATED TRANSMITTER


FREQUENCY RESPONSE OF TXV 13 TRANSMITTER AT MODULATION WHICH PRODUCES
●● NO VISIBLE DISTORTION OVER USEABLE FREQUENCY RANGE ●●
3
2
1
OUTPUT
AT
RECEIVER
VOLTS
WITH LOW FREQUENCY
FM MODIFICATION
.8 VOLT RUNS INPUT TO TRANSMITTER
CYCLES
1 Kc
3 Kc
10 Kc
30 Kc
100 Kc
300 Kc
1 Mc

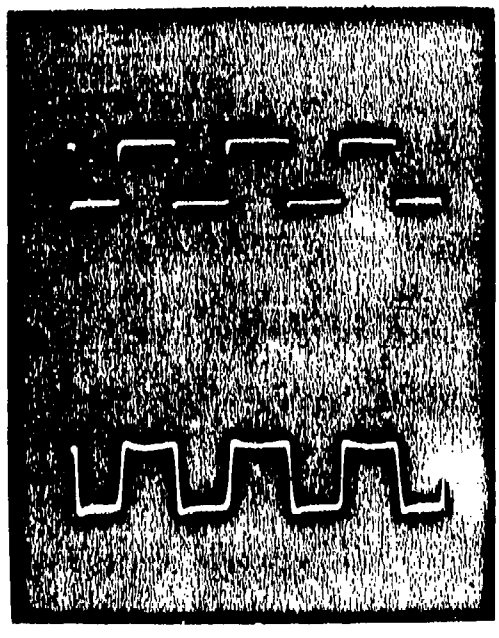

RELATIVE GAIN DB
0
1
2
3
4
5
6
7
8
9
10
DASHED CURVE
INPUT CAPACITOR - 1 MFD
MOD. GRID RESISTOR 100K
SOLID CURVE
STANDARD XMTR
FREQUENCY RESPONSE OF FM
TM TRANSMITTER
SIG. GEN HP 650A
TXV 48
CLARK mod. 167-E
SCOPE 514 D
100
1,000
10,000
100,000
FREQUENCY CPS



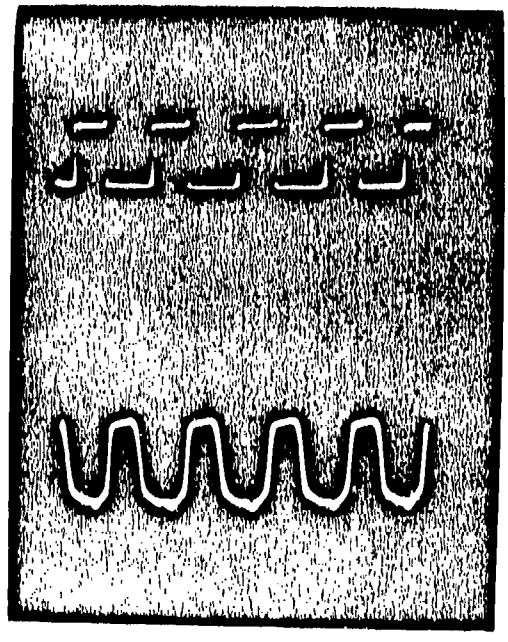
1KC



5KC

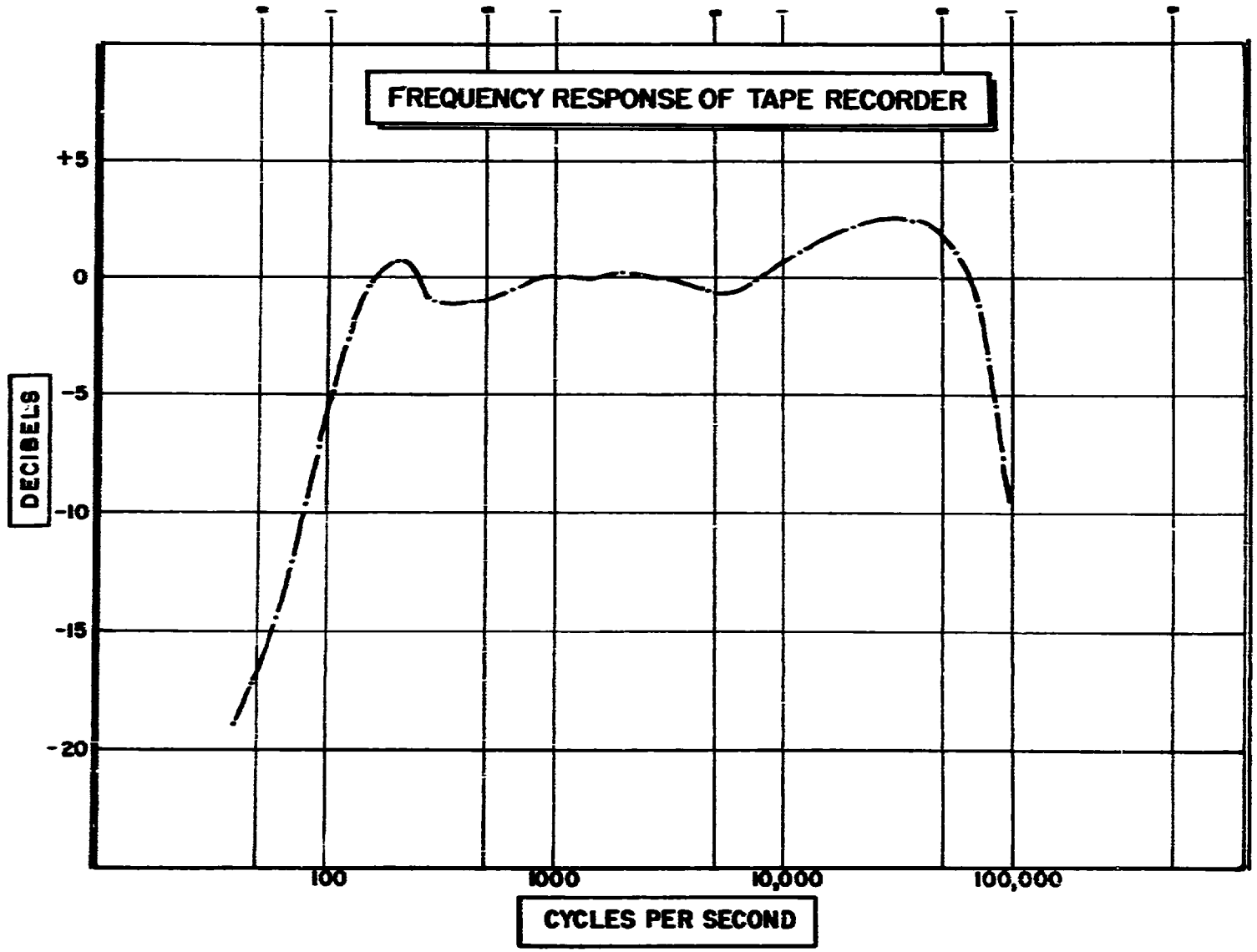


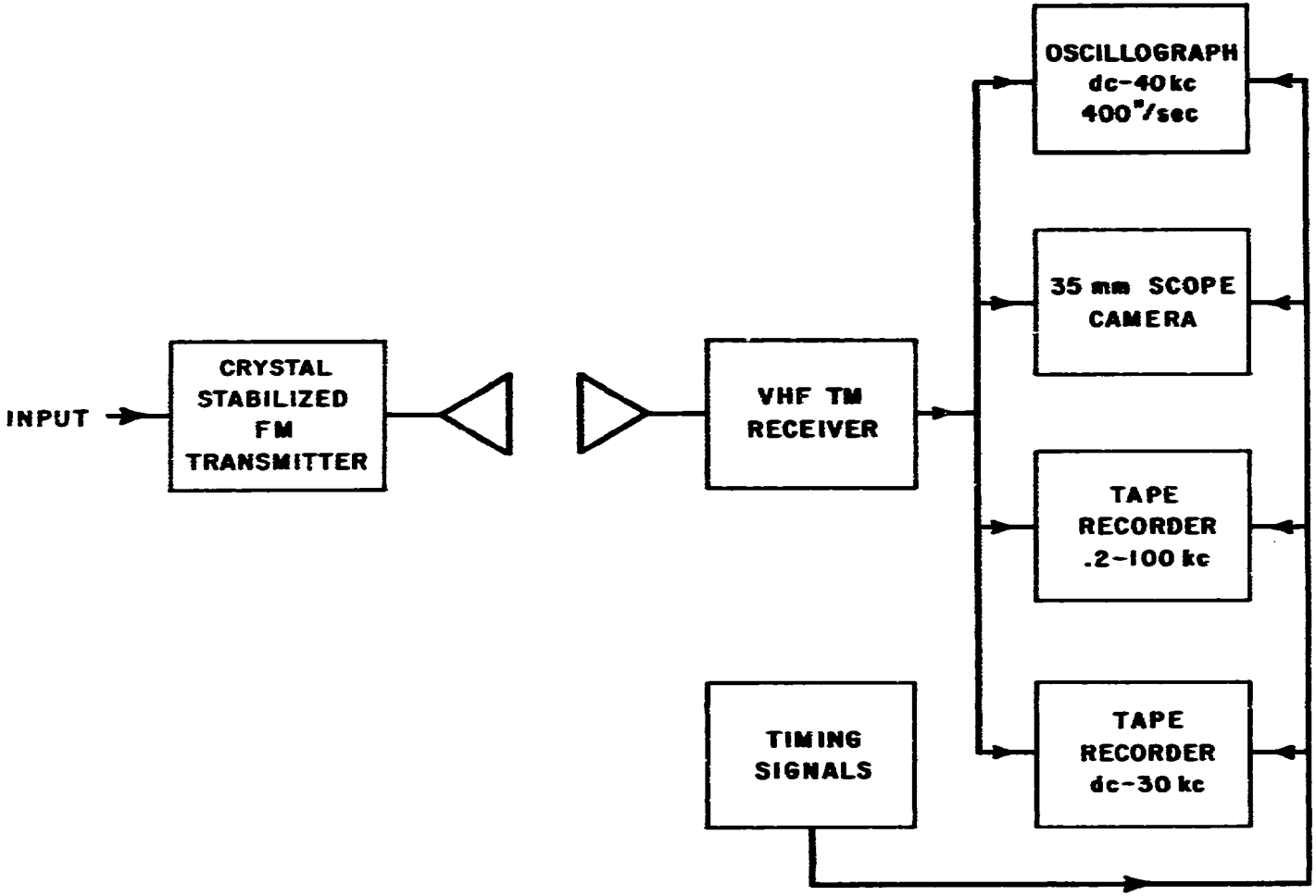
10KC



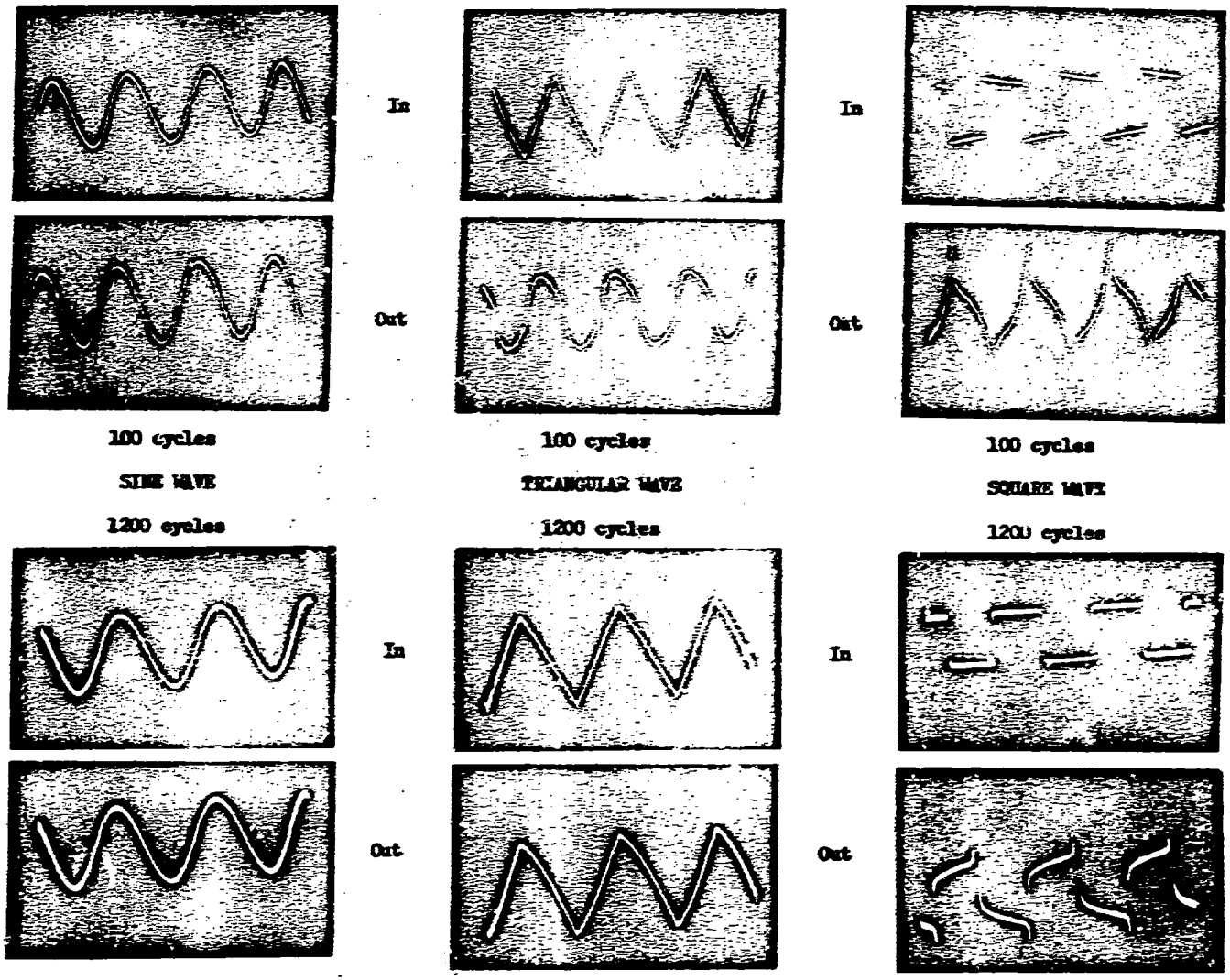
20KC

SQUARE WAVE RESPONSE OF FREQUENCY MODULATED TRANSMITTER


FREQUENCY RESPONSE OF TAPE RECORDER
+5
0
-5
-10
-15
-20
DECIBELS
100
1000
10,000
100,000
CYCLES PER SECOND



BLOCK DIAGRAM OF WIDE BAND TELEMETERING SYSTEM



WAVE SHAPE REPRODUCTION CAPABILITIES OF TAPE RECORDER

AUTOMATION - WHERE IS THE POINT OF DIMINISHING RETURN

BenAmi Blau
Human Engineering Laboratory*
Aberdeen Proving Ground, Md.

In choosing a topic for discussion, I sought a subject which I felt would be of timely interest, relatively clear cut, and generally without a great deal of controversy about it. From a layman's point of view, the topic of automation seemed to fill the bill. I was extremely naive in my choice. Fortunately I chose to deliver this presentation in a clinical session of this conference. I do so utilizing that definition of clinical session which allows the presentation of a problem area with no answers or solutions required from the speaker.

My interest in automation is in the man-machine integration involved in a complex system. Contrary to the layman's popular conception of automation, which is essentially the pushing of a start and stop button in response to a red or green light, there may be a more intricate relationship involved. With this thought in mind, I diligently began what I intended to be an intensive literature review.

Much of the published literature on automation is concerned with semantic arguments of definition, economics, and the pros and cons of the effects of either a benefactor or monster on society, depending upon whether management or labor was speaking. Only one point of commonality appeared to exist. The major portion of the material in the literature began with a definition. The definitions were varied and not always in total agreement. As examples, some of the definitions were:

1. Automation means automatic control. (1)

2. The substitution of mechanical, pneumatic, hydraulic, electrical and electronic devices for human organs of decision and effort.

3. The science and art of manufacturing products with minimum labor, effort and cost, and maximum efficiency.

4. The elimination of repetitive, onerous, dangerous and trivial labor, mental or physical, from the realm of human endeavor.

5. The way to a society in which labor is necessary not for the physical needs of the body, but for the creative needs of the soul.

In a final definition for purposes of this discussion, a distinction is made between mechanization, which replaces or amplifies human brawn, and automation, which supplements the human brain through the inclusion of feedbacks or self-correcting devices. (3).

Following the established pattern of defining the term, an operational

* After 15 February 1957 the author will be associated with the International Business Machine Corporation (Endicott, New York).

definition for purposes of discussion is proposed:

> Automation - the substitution of a mechanical and/or electronic device in a man-machine system for a function previously requiring human perceptual, cognative, memory, decision making capabilities or psychomotor response.

This paper's interest in automation is from a man-mahcine integration standpoint. The type of automated systems of prime interest are weapon systems. These fall into the category of fire control and guidance systems, primarily for guided missiles. Automation, using the operational definition, is an inherent component of all of these systems. The question to be raised at this time is how far should these systems be automatized considering the reliability of the output of the entire system. Since it is not probable that we will have systems entirely independent of human influences in the immediate future, either from an operational or a maintenance standpoint, we will still be dealing with man-machine systems. There are several factors which must be considered in the establishment of criteria for a point of diminishing returns in an automated system. These criteria are factors which strongly influence the reliability of the total system. Perhaps at this point the term reliability as used here should be defined. The term "reliability" should mean essentially the probability of a man-machine system, in this case a weapon system, accomplishing its military mission.

Factors inherent in the system which will influence the reliability of the system are:

1. Complexity of the mechanical, electronic, hydraulic and communicative components of the system.

2. Reliability of the parts making up these components.

3. Environmental factors such as temperature extremes, vibration, shock and accelleration which will influence the reliability of both parts and components.

4. The quality and quantity of manpower required to operate and service the system.

5. Environmental and mental stresses placed on the manpower serving the system.

Further consideration must also be made as to the intended use, from a tactical point of view, of any particular weapon system. Requirements exist which limit the size and weight of weapon systems. Consideration as to production cost, maintainability, and transportation of such systems must also be made. How then may criteria be established which will provide the planners and designers of automated equipment with sufficient information so as to enable them to develop systems which will meet both technical and tactical specifications. A fairly obvious answer presents itself immediately. Merely determine the capability and reliability of functioning of the particular machines involved and the capability and reliability of the men who

must serve this equipment.

The answer sounds fairly simple. The implementation, however, leads us into a variety of problems. One of the aims in the design of electronic machines is the development of high performance equipment using automatic control, guidance and computing features. The incorporation of these features generally results in more complex and sometimes less reliable equipment. A serious question previously raised by Boodman (2) is the problem of deciding what degree of reliability in a given operation is acceptable and of determining the degree of complexity in a machine that will decrease the reliability beyond this acceptable value.

To consider the dependence of equipment reliability upon equipment complexity, the factors previously mentioned as affecting component reliability must be understood and a measure of equipment complexity must be established. By the same token, the complexity and reliability of the human component of a man-machine system must also be considered. Factors affecting the reliability of a human operator in a broad sense, are somewhat similar to those factors which affect the reliability of machines. Environmental extremes, visual limitations, auditory limitations, noise and vibration are some of the external forces influencing the human being. Unfortunately, the human machine cannot be subjected to standardization of parts and quality controls in production. Therefore, we must consider individual differences as factors of fatigue, motivation, perceptual stress, cognitive ability and psychomotor limitations in determing the reliability of the human being. The problem of establishing reliability criteria on either machines or man is in itself a most difficult task. The problem becomes even more difficult when the relationship between the complexity and reliability of machines and the functioning of human components within the system must be combined to arrive at an output figure.

No definitive or inclusive approach is known to the writer at the present time. It is intended that this problem will stimulate the thinking of scientists concerned with the design of complex machines and the people who must operate and maintain them in order that hypotheses be formulated and tested which will lead towards even partial answers to the problems.

One approach which may be considered is that of attempting to establish a quantitative and qualitative measure of complexity for machines. In these measures of complexity must be incorporated varying degrees of necessary human input. Perhaps by studying the interaction of man and machine in a variety of situations which range from simple to complex in terms of both human and equipment functioning, criteria may be developed which will indicate fixed points delineating optimum functions in integrating man and machine.

DETERMINING WHETHER A PRODUCT MEETS TASTE SPECIFICATIONS

Norman J. Gutman
QM Food and Container Institute
Chicago, Illinois

The quality of many food products cannot be determined exclusively by objective physical and chemical tests. Thus, it is necessary to have some measure of taste or palatability by a consumer or expert panel. Therefore, the specifications written for various foods require that certain taste or palatability criteria be met. Our problem is that of setting these criteria on a basis which protects the legitimate interest of both the government and the producer.

In practice the problem arises in two separate stages. First, a standard for a satisfactory product is to be established. Second, as individual contracts are let, it is necessary, by pre-award testing, to determine whether the product submitted for evaluation meets the established standard.

In the first stage, establishing a standard, the present practice is to have a group of persons, either military or civilians, at an Army post or at the QM Food and Container Institute, all depending on the particular product, rate certain samples on nine point scales. The standard most commonly used is a preference scale called the hedonic scale; a quality grading scale is somewhat less frequently used. This talk will not consider the questions of adequacy of scale, dimensions of preference, effect of one sample upon the rating of another, and other such problems discussed by Professor Bradley. Whether these limitations will lead to a serious oversimplification is a point which might well be considered.

Most specifications, as presently written, require that any sample whose mean scale rating is significantly below the mean rating of all samples at the 5% level be rejected from further consideration in establishing the standard. Depending on the specification, the test may range from the erroneous application of a multiple Student t test using the gross variance of the rating of a sample to a multiple range or multiple F test such as those developed by Duncan, Tukey, Dunnett, or Bechhofer. However, none of these tests is directly valid since ratings for any one sample must be compared with the average rating of all samples.

A question which immediately arises is why a sample's rating should be compared with the average rating of all samples. The justification essentially is that the samples submitted for testing are representative of the quality of product available, and only those samples which rate sufficiently below the average should be rejected. It may be added that markedly inferior samples are usually screened out by chemical and physical tests before the taste test is run.

The second stage arises after standards for the product have been established. Now the problem is to see that the samples of product submitted for pre-award evaluation meet these standards. But in this case, is a product to be compared with its previous quality or with the standard of satisfactory products established at the previous evaluation? The

latter comparison has been preferred since in the first, if a product is of very high quality in the first evaluation but is of lower quality on the second, it is rejected while another product of the same quality in the second evaluation but of lower quality on the first will be accepted.

Here the problem of comparison with the standard obtained in the first evaluation arises. One might compare the average level of ratings for satisfactory products in the first evaluation with the average rating for a product in the second (or pre-award) evaluation. Since it is known that the level of preference ratings may vary considerably with time and with the group rating, this comparison is somewhat untrustworthy. Thus, if the general level of ratings on the first test is high, while on the second test it is low, pre-award samples may be rejected even though they are of as good a quality as those on the first evaluation. Conversely, a low rating group on the first evaluation and high rating group on the second evaluation may result in a poor quality product's being purchased.

In the past it was necessary to follow this practice in all products, and it is still followed in some products. In a few products whose quality is not affected too seriously by a reasonable length of storage (say one year or less) satisfactory samples of the products from the first evaluation are stored. Then, when a pre-award evaluation is necessary, samples from the previously satisfactory productions are tested along with the pre-award samples; thus a more legitimate comparison can be made. In other products where manufacturing practices permit, a different method is used. Through procurement or its own production, the Institute obtains satisfactory samples of a product to be established as a standard. Then these standard samples are submitted by invitation to a group of producers who are asked to submit pre-award samples at least as good as the standard sample. Then on pre-award evaluation, the pre-award samples are tested together with the standard. However, some products are relatively perishable, and so those procedures cannot be followed, and the pre-award samples must be compared with the average level of ratings of satisfactory products in the first evaluation.

In the first two situations where a direct comparison among the pre-award samples and the standards can be made, a test such as that of Dunnett (JASA, Dec 1955) is readily applicable. In the remaining situation, where the pre-award samples are compared with the level of ratings set as the standard, the specifications as presently written require that a multiple t test be used. It appears that the Duncan, Tukey, Dunnett, and Bechhofer multiple comparison tests are not directly applicable to this problem.

These are problems which vitally affect the Armed Forces, and any assistance in their solution will be deeply appreciated.

EXPERIMENTAL DESIGN FOR DETERMINING SPECIFICATION LIMITS FOR MANGANESE-ALUMINUM BRONZE

S. L. Eisler
Rock Island Arsenal

Federal Specification QQ-C-523 covers the procurement of manganese and manganese-aluminum bronze ingots for remelting. There are several alloys with various chemical composition limits specified. In addition, there are mechanical property requirements such as tensile strength, yield strength and elongation.

The problem which has been encountered on numerous occasions is that suppliers are able to easily meet the chemical requirements, but not the physical requirements. This naturally leads to a great deal of discussion as many suppliers feel that if the material passes the chemical analysis it will possess the mechanical properties required. Unfortunately, this is not true and it is the opinion of the metallurgists at Rock Island Arsenal that the limits for chemical composition are too broad. It is also their opinion that conditions of preparation of the ingots, although contributory to their final properties, are of minor significance. Therefore, we are interested in studying the changes in physical properties as the percentage of each alloying element is varied within the specification limits.

For example, let us consider the requirements for Alloys B & C which has the same chemical composition limits but different mechanical property requirements:

Chemical Composition

Copper	60 - 68%
Aluminum	3.0 - 7.5%
Manganese	2.5 - 5.0%
Iron	2.0 - 4.0%
Tin	$<$ 0.10%
Lead	$<$ 0.10%
Nickel	$<$ 1.0%
Zinc	Remainder

Mechanical Properties

	B	C
Tensile Strength (min. psi)	90,000	111,000
Yield Strength (min. psi)	45,000	60,000
Elongation (min.)	18%	12%

This particular example presents a more complicated problem due to the dual set of mechanical property requirements. However, it has been chosen as an example as it is believed that separate chemical compositions should possibly be specified for each alloy. Although this example is more complicated than the other alloys specified, the same difficulties have been encountered.

The problem which we would like to present to this clinical session today is how can we design an experiment to determine reduced limits for the more important elements, such as copper and zinc, which will insure conformance with the mechanical requirements. It will be noted that as the copper content is increased the zinc content is similarly reduced, providing that the contents of the other elements are unchanged. This presents a difficult situation as you can not change the content of one element independently of the other.

It has been suggested that an extensive review of past data and comparison of the composition and mechanical properties of past lots might prove valuable. However, after checking over some of the past data it was found that insufficient information was available.

Therefore, we are open for ideas which will simplify this investigation. Perhaps someone present has encountered a similar metallurgical problem.

I might add that this problem is not common to this specification alone. It is also quite common to Federal Specification QQ-B-675 which covers Aluminum-Bronze Ingots.

SAMPLING PLAN FOR PACKAGING MATERIALS PRODUCED BY A CONTINUOUS PROCESS

S. L. Eisler
Rock Island Arsenal

The Department of the Army purchases a large number of packaging materials which are products of a continuous manufacturing process. This is true of various paper products, barrier materials, textiles, tapes, etc. During the manufacturing process, the continuously produced product is rolled into convenient sized rolls. Unfortunately, in most cases the identification of rolls in order of production within a lot is not available.

Thus, an inspector may be faced with the problem of selecting a representative sample from a shipment of 100 or more rolls for laboratory tests. There are numerous sampling methods presented in the literature for sampling carloads of coal or salt, tank cars of oil or acid, and, of course, the numerous methods of selecting a sample of a discrete manufactured unit. However, the problem mentioned above is unlike any of these situations due to the fact that samples from the interior of the rolls are not readily accessible.

Therefore, it is believed that the first step must be a study to determine the magnitude of the various sources of variability. The three major sources of variability are probably:

1. Edge to edge variation.
2. Within roll variation.
3. Between roll variation.

From the results of this preliminary investigation conducted on products from a representative cross-section of suppliers, it should be possible to test the significance of the variabilities of the above three sources against the variabilities of the different tests employed.

Based on the above comparisons, definite sampling plan recommendations could be made which would result in samples which would reflect the variations considered significant. For example, if the edge to edge variation were the only one found to be significant, one sample taken from any roll would be sufficient, provided the individual test specimens were randomly chosen from the sample.

Many of the current military specifications for materials of this type state the sample size, e.g., in square yards, and even specify the number of square yards to be taken from a roll. However, it is believed that these choices have been made without a realistic statistical evaluation of the material, such as is proposed.

There are now two or three questions I should like to present to the panel and the others in attendance.

1. Does our approach to the problem appear to be reasonable?

2. Does anyone know of any similar materials which have been studied? If so, what type of sampling plan resulted from these studies?

3. How may the various procedures for sampling inspection by variables (ORD-M608-10) where definite units of product are specified be converted to apply to material produced by a continuous process? What constitutes a unit of product for material of this type?

EXAMPLE

For example, ORD-M608-10 specifies for a lot size of 10,000 (assuming we have 10,000 sq. yds. of material in the lot and have designated 1 sq. yd. as the unit of product) a sample size of seventy. MIL-B-121A, which covers barrier material, specifies 40 sq. yds. for a similar size lot. The total amount of material required for the laboratory tests is approximately 6 sq. yds.

Now, the question arises as to how the test specimens are to be distributed throughout the sample. There is also no way in which the acceptance criteria of a variables sampling plan may be used where a measurement is not made on each unit of product but where a number of measurements are taken on the entire sample made up of several units of product.

OBSERVATION ON THE USE OF MODELS IN THE DESIGN OF EXPERIMENT

James W. Mitchell
Frankford Arsenal

The importance of models in statistics is almost obvious. Mathematical models are widely used to express statistical tests and as a basis for deriving new ones. However, exact mathematical models in equation form are usually no easier to understand by scientists and engineers in other fields than the rest of the language of statistics. In communication between statistician and other scientists and administrators, models can play an important roll in clarifying understanding of a problem in statistics.

One would begin by statement of the hypotheses in terms of models - but not necessarily mathematical forms. Thus the problem is defined in a form understood by the statistician as the bases of a well defined statistical test and by the engineer as a form which his collected data may take. It should therefore be of tremendous help in refining the statement of the problem to the mutual satisfaction of both statistician and scientist and thus form a common meeting ground for the two. It is my thesis to try to exploit this property of models to greater advantage to improve the communication between scientist, engineer and statistician.

A discussion of scientific models can lead one far into the field of philosophy and logic. This would be unwise to attempt. However, it is well to recognize three levels of model making. First, a complete scientific model of an experiment would encompass all the possible concepts and relations which a scientist could use and thus it is an ideal of science. It could involve the whole wealth of modern logic and mathematics, the fields of science needed to describe the possible phenomena and the definition involved thus requires the aid of psychologist and sociologist as well. .It is quite a formal structure and probably never has been fully realized in any field. Today many partial models are being constructed to suit the various sciences. The expression of physical laws or the statistical concepts which we have been hearing about in terms of mathematical equations represents these partial models. These are the working models used by the scientist in his field to advance his study of the science. However there is still another level of models needed today. These are models required to create common understandings between dependent but different fields of science on the level of the common worker.

Let's examine an example of the formation of a problem model. One observes a difference in some measured property between two or more groups of items and forms an explanation of the difference. This explanation is then contrasted with the universally applicable hypothesis of randomness. Statistics are then applied by creating a specific statistical (null) hypothesis or model out of the vague concepts of random phenomena. Sometimes the choice of a statistical or random model is obvious; sometimes it is far from easy to find an acceptable model to match the natural situation. A model may also be devised for the alternative hypothesis corresponding to the physical explanation of the difference. Although it is often not needed, the latter would be essentially one of difference,

correlation or non-randomness. The statistical test is then applied by comparing the experimental data, collected under the assumption of random sampling, with the statistical model. A choice between the null and alternate hypothesis is then made according to whether or not the composition of the data can be explained by this statistical model. The model must be specific in the sense that one can calculate from it the probability of occurrence of deviations from the assumed average composition of the model. The magnitude of the deviation of the experimental data from the assumed statistical model then forms the basis of choice between the null and alternate hypothesis, i. e., between the statistical and physical models of the experiment.

Statistical procedures which fit the above example are the comparison of two or a set of averages or variances and related tests. The random model for these is the normal distribution. This model is easily understood and can be concretely illustrated in a variety of ways (e.g. the Quincunx). Other closely related models are the binomial and Poisson distributions. The familiar urn containing balls of two colors is a physical model of these distributions.

In order to form a logical basis of the statistical test the model should have certain properties. These are: first the property of being specific in the sense that it permits the adoption of specific statistical procedures. The normal distribution is a good example of this. Models must also satisfy certain requirements of randomness and may contain arbitrary elements that are not "natural" but which do not conflict with the possible alternate hypotheses.

It is certainly not necessary to construct a model about the null or statistical hypothesis. Physical concepts which can be expressed in mathematical form are best represented by this "mathematical model". The physical concept usually implies causality. The simplest form of a mathematical model would probably be a linear regression in two variables. In a more general example there are multivariate regression, power functions and any number of possible mathematical forms representing specific types of causality and even natural law. In each case the a' priorie assumption of one of these relationships constitutes a mathematical model of the portion of the physical universe to be examined. One then wishes to see if the experimental data are consistent with or will support this hypothesis. The procedure of statistical testing requires the creation of an alternate statistical or random model in which the display of experimental observations are attributed to chance alone. These statistical models are usually more complicated than the simple normal, or other distributions referred to previously. In fact the statistical model can be considered as N dimensional for an N dimensional physical law. However to be able to treat the results quantitatively with the usual tests of significance, some specific distribution function must be assumed and applied one dimension at a time, i.e., coefficient by coefficient. The mathematical and the statistical model may then be used together to illustrate the application of statistics to the problem.

Another class of models are those on which the factorial experiment and randomized block designs are based. The statistical model is a randomized area, or N-dimensional volume as for the causal relationship above and the physical model is a form of the multivariate equation but which includes terms for experimental error and other variation as well as the main variable terms and their interactions.

I hope that these examples are sufficient to illustrate some forms that a model may assume. These may be mathematical, physical such as the urn and balls or a roulette wheel, spatial as an N-dimensional model or even mechanical such as a model to show the interaction of tolerances. The model is a type of model which is comprehended by the engineer in some familiar dimensional or spatial form and by the statistician as a specific model of a random distribution of objects or events is a particularly useful form for improving the experimental design. If these two start by reducing the problem to a statistical model of the null hypothesis, the similiarity of this model to the preconceived physical or "natural" model of the experiment will be easier to see. The statistical and physical models can then be refined until the experimenter is satisfied. The physical model, thus defined, becomes an alternate hypothesis and this interplay may even lead to other alternate hypotheses that deserve consideration. Often it may happen that a scientist is led to accept one statistical procedure as best suited to his need when it is not entirely appropriate to his experiment. The practice of first settling on the correct model with several possible statistical tests in mind should prevent this and would permit full utilization of the model as a joisting ground between the experimenter and the statistical until an acceptable test is found.

SHORT RANGE SCATTER PROPAGATION SURVEY

Messrs. Lacy, Sharp and Lindner
Signal Corps Engineering Laboratories

INTRODUCTION: The technique of photographing the returned terrain scattered power, observed on a radar scope, and then overlaying the photographed terrain scattering areas on a properly oriented contour map of the swept area surrounding the radar location, displays immediately the radio line-of-sight paths. Such displays of the scattering areas indicate all prospective communication paths between the location of the radar and the areas producing the scatter. The returned power from the scattering areas shown on the contour map must be correlated with the system gain of the microwave communication equipment to be employed. Information relative to the actual path transmission loss between the location of the radar and any point in those areas producing the scatter, would definitely determine the feasibility of a prospective communication site. This information is not obtainable from the photograph and would necessitate an actual path transmission loss measurement between the location of the radar and the particular point in the areas producing the scatter. This is not feasible for the intended application of the above mapping technique for the siting of short range microwave communication equipment with fifteen foot high antennas.

DISCUSSION: The actual path transmission loss for the microwave frequency to be used is the sum of the free space path transmission loss determined by the distance between the radar location and the proposed communication site, and the terrain factor power loss determined by the type of terrain along the communication path. There is not available to date sufficient data that would correlate the type of terrain of the communication path with the terrain factor power loss. Were such a correlation available, then, from such a map overlay as shown in Fig. 1* and with a knowledge of the type of terrain of the communication path, the feasibility of establishing communication over the path involved could be readily determined. We are now concerned with the obtaining of such data and the best means of establishing such a correlation, if it exists, from the experimental data obtained to date. This, it can be readily seen, will not be an easy task when one considers the many types of terrain that can be involved and the magnitude of the contribution to the communication path transmission loss of the terrain factor, particulary as it is affected by the terrain in the immediate vicinity of the transmitting and receiving sites.

In addition to the photographing of the scatter pattern observable on the radar scope, the received scattered pulse amplitude from the area at the desired communications site is compared to the radar transmitted pulse amplitude. The received amplitude scattered by a given area is from many scatterers comprising the area. The received amplitude from the scattering area is compared to the radar transmitted amplitude. From a measurement the ratio of the receiver input scattered return average power to the radar transmitter average power output expressed in db is obtained.

* Figures appear at end of the article.

Goldstein has shown in the book "Propagation of Short Radio Waves" of the MIT Series that the total average radar received scattered signal power summed up for many scatterers in the target area, where the same antenna is employed for transmitting and receiving, is the following designated equation (1).

$$(1)\quad \overline{P_R} = G^2 \sum_j \left\{ f^4(\theta_j, \phi_j) \left(\frac{3\lambda}{8\pi R_j} \right)^2 \left(\frac{\sigma_j}{4\pi R_j^2} \right) \right\} \left\{ P\left(t_o - \frac{2R_j}{c} \right) \right\}.$$

In this expression, G is the maximum antenna power gain, the first factor under the summation sign is the antenna pattern function, the second factor is the free space power loss referred to a doublet radiator, the third factor is a measure of the scattered power as a function of σ_j - the scatter cross section of the "jth" scatterer, and the last factor under the summation sign is proportional to the magnitude of the Poynting vector of the incident wave at R_j at such a time that the reflected echo from the "jth" scatterer returns to the radar at the instant of time t_o. At those distances from the radar to the target area where it can be assumed that the sum in equation (1) involves a very large number of scatterers, the summation may be replaced by tn integral where

$$\left\{ N\ (R, \phi, \sigma) \right\} R\, dR\ d\phi\ d\bar{\sigma}$$

is the density function which gives the number of scatterers in an area element R dR dϕ for which the radar cross section lies between $\bar{\sigma}$ and $\bar{\sigma} + d\bar{\sigma}$, and where it can be further assumed that the scatterers are distributed uniformly and homogenously over the target area so that the function N is only a function of $\bar{\sigma}$ and the free space power loss is independent of R_j, then equation (1) becomes equation (2).

$$(2)\quad \overline{P_R} = \left(\frac{3\lambda}{8\pi R} \right)^2 \left[G^2 \int_{-\pi}^{\pi} \left\{ f^4(\theta, \phi) \right\} d\phi \right] \left[\frac{\int_o^{\infty} \left\{ N(\sigma) \right\} d\bar{\sigma}}{4\pi R} \right] \left[\int_o^{\infty} \left\{ P\left(t_o - \frac{2R}{c} \right) \right\} dR \right]$$

In the expression designated as equation (2), $\bar{P}_R$ is the total average radar received scattered power. The first factor on the right is the free space path power loss, the second factor is the combined power gain of the transmitting and receiving antenna modified by the antenna pattern over the target area, the third factor is a measure of the scattered power from the target area, and the fourth factor is the transmitter power output. Now equation (2) holds approximately for distances in excess of six miles. As the distances increase, the more accurate equation (2) becomes. For distances less than six miles equation (2) does not hold, and equation (1) involving the summation from individual scatterers must be employed. If the radar transmission path is over a terrain, then a two-way terrain loss factor must be added.

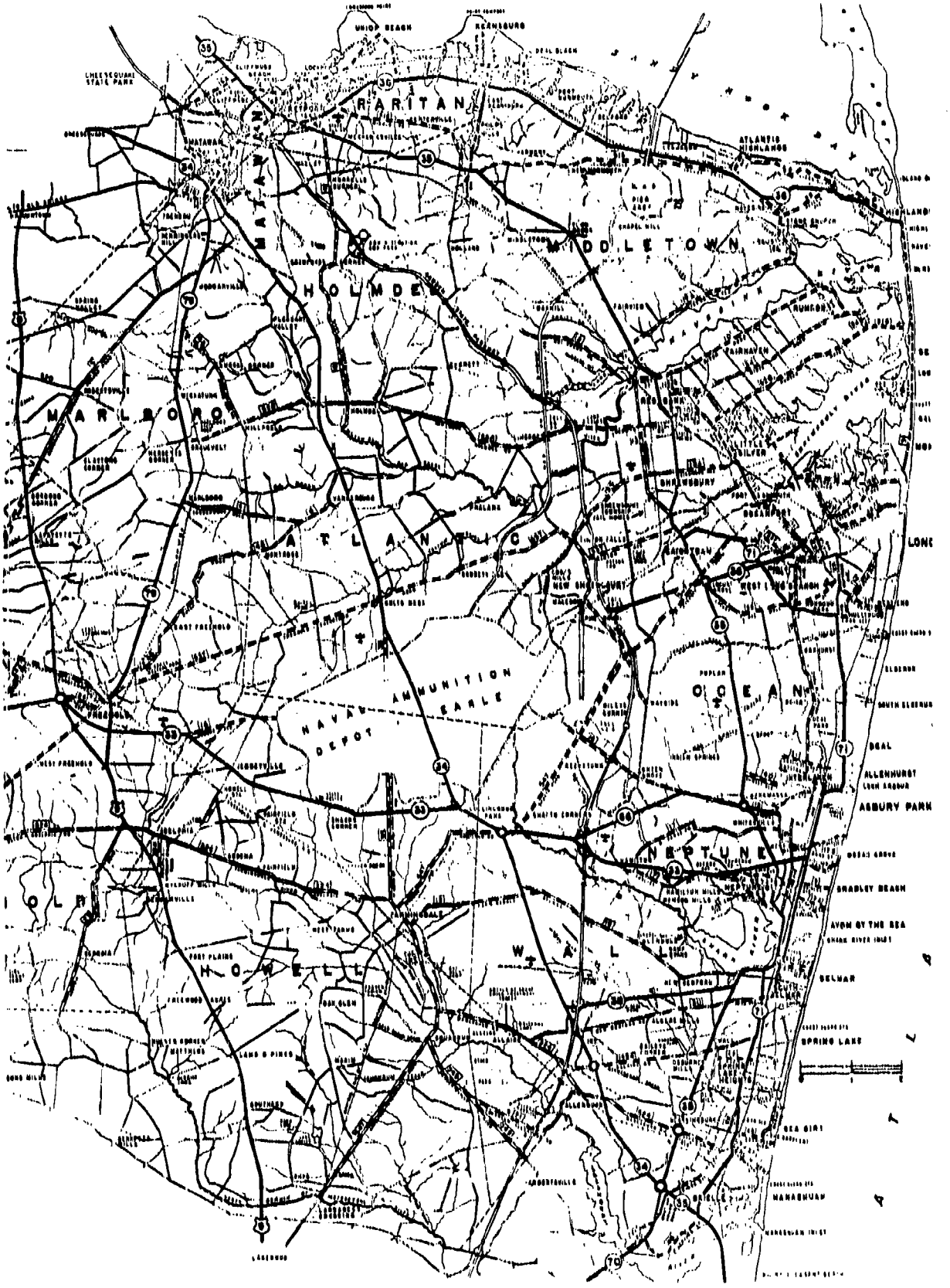
Hence for distances in excess of six miles, it is approximately accurate that the ratio of the radar receiver input target area scattered

power to the radar transmitter power output expressed in db is equal to combined power gain of the transmitting and receiving antenna modified by the antenna pattern over the target area expressed in db, plus the free space power loss expressed in db for the distance from the radar set to the target area, plus the terrain factor power loss expressed in db for the transmission path from the radar set to the target area and return path to the radar set, plus a loss expressed in db which is a measure of the scattered power from a selected target area. That is in equation (3).

$$(3)\quad 10 \text{ Log } \frac{P_R}{P_T} = 10 \text{ Log} \left[G^2 \int_{\pi}^{\pi} \left\{ f^4(\theta, \phi) \right\} d\phi \right] + 10 \text{ Log } A_o^2 + 10 \text{ Log } A_p^4 + 10 \text{ Log } \mathfrak{X}_R$$

P_T is the radar transmitter average power output, P_R is the radar receiver average input target scattered power. The first term on the right is the combined power gain of the transmitting and receiving antennas modified by the antenna pattern over the target area expressed in db. The second term is the free space power loss referred to a doublet radiator expressed in db. The third term is the forward and return terrain factor power loss expressed in db. The fourth term is a measure of the target area scattered power expressed in db.

If a correlation can be obtained between the communication path terrain power factor loss and the various types of terrain of the communication paths, then with the aid of the radar power ratio measurement a correlation between the type of terrain along the communication path and the radar power ratio measurement may be obtainable. Fig. 2 is an example of how the experimental data is presently summarized. In the twelve rows are the results of twelve field measurements. Column 1 is the path length in miles; column 2 is the calculated free space path transmission loss A_o^2 (Eq. 3); column 3 is the measured terrain loss factor (the measured path loss from column 4 minus the free space loss, column 2). Column 5 is the ratio of power level received to power level transmitted by radar; in column 6, $\mathfrak{X}_R$ is a measure of scattered power from the target area. Column 7 is for the terrain type classification. The problem submitted is the need for method of terrain classification that will permit a predetermination of the path transmission loss from the physical aspects of the terrain. Column 8 is the relative heights of the selected sites, another factor believed to be an important consideration for the prediction of the terrain loss factor.


RARITAN
MIDDLETOWN
HOLMDEL
MARLBORO
ATLANTIC
OCEAN
NAVAL AMMUNITION DEPOT EARLE
NEPTUNE
WALL
HOWELL
LONG BRANCH
ASBURY PARK
BRADLEY BEACH
AVON BY THE SEA
BELMAR
SPRING LAKE
MANASQUAN

1.

	1	2	3	4	5	6	7	8
	Path Length	Free Space A_o^2	Terrain Loss A_p	Path Loss	Power Ratio $\frac{P_r}{P_t}$	Scatter Power X_r	Terrain Type	Relative Elevation
1	6.00mi.	111db.	26db.	134db.	139db.	-30db.		+ 6ft.
2	6.75	112	21	133	121	-23		- 85
3	8.25	113	5	118	140	-72		+ 40
4	8.25	113	23	136	133	-29		+ 40
5	8.25	113	26	139	145	-35		- 90
6	8.75	114	23	137	139	-34		- 50
7	8.75	114	24	138	131	-24		+ 40
8	8.75	114	23	137	139	-34		0
9	9.00	115	41	156	144	- 3		- 10
10	9.00	115	16	131	130	-39		- 10
11	9.00	115	23	138	117	- 2		+ 110
12	10.00	116	32	148	139	-15		+ 8

FIG. 2

EXPERIMENTAL DESIGNS FOR ORGANIZATION RESEARCH USING LIMITED RESOURCES

Raymond H. Burros
Combat Operations Research Group
Ft. Monroe, Va.

The title of this paper is somewhat misleading, since I do not intend to discuss specific details of experimental design. Instead I shall first present a methodological problem growing out of limitations in resources available for experimental research in military organization. Then I shall present some possible approaches to the solution of the problem without going into details of experimental design. In a sense, therefore, the discussion will deal with classes of designs. Some of the more crucial assumptions will be examined. I shall conclude the discussion by presenting a possible approach which may lead some of you into some new lines of thinking.

SOME CHARACTERISTICS OF ORGANIZATION RESEARCH

Research in human organization has at least two important characteristics distinguishing it from research on individual organisms, human or otherwise. First, the experimental unit is not the single human being; it is a specified kind of human group, such as an infantry platoon. Second, the group score is frequently obtained by observing the behavior of the group, but not necessarily the detailed behavior of each member of the group. In other words, the group score is often not simply the sum or mean of the scores of the members of the group, although these members help to determine the group score.

These characteristics imply that a fairly large number of subjects is needed to gather data on a relatively small number of types of organization. The limitation on number of troops available for use as subjects is most pressing. Other types of limited resources include terrain and equipment.

To make the problem more concrete, let us make some specific assumptions. First, we have available a regimental combat team, i.e., the equivalent of an infantry regiment with additional supporting weapons units. This will provide 27 rifle platoons of the present size with other weapons units. Second, different sizes and structures of the infantry platoon provide the independent variables, and various measures of effectiveness are the dependent variables. Third, some of the experimental organizations will demand more enlisted men than does the present day platoon. The problem is to choose an approach to experimental design which will take account of resource limitations and still be powerful enough to detect reasonably important differences.

POSSIBLE APPROACHES TO SOLUTION

The first approach is to assign at random some of the 27 existing platoons to the various treatments (organization structures). The members of the remaining platoons are used to augment those platoons which require more than the presently allocated strength. This gives us

(say) a total sample of twenty experimental platoons. Because of the great variability of group scores, however, this approach is probably not powerful enough to detect differences between as few as four types or organization.

The second approach is to use depleted or "skeletonized" military units. These would have full complements of commissioned and non-commissioned officers but would simulate the existence of most of the enlisted men. Although there may be some possibility of doing this, it would still be necessary to validate the methodology by means of experiments with complete military units. Therefore, this approach does not solve the full problem.

The third approach is to take a number of platoons and run each under all of the treatments, when the number of these is small. This approach assumes that an existing platoon preserves its essential identity even though noncommissioned officers and enlisted men are randomly added to it or removed from it to fit the structure prescribed by the experimental treatment. Then each of twenty existing platoons can be run under all of the treatments when the number of treatments is small.

An adequate design for this approach will have to control for two kinds of order: first, the order in which the treatments are applied to each platoon, and second, the order in which the platoons are tested. If there is no reason to expect that either order interacts with treatment, then several kinds of experimental designs can be applied. There is good reason, however, to expect interaction between treatments and the orders in which they are applied to the platoons. Presumably once a platoon has learned to function under one organization, this learning may either facilitate or inhibit its performance under a different organization. Psychologists recognize this as the process of positive or negative transfer. Our knowledge of transfer is not adequate enough to predict exactly what will happen. It is sufficient, however, to justify my assertion that interaction is likely to be both present and large. If this is so, then such an approach will not yield trustworthy conclusions about the relative effectiveness of different kinds of organization. Although I do not mean to assert that this approach of applying all treatments to each platoon is hopeless, it may be worthwhile to consider another approach.

The fourth and last approach to be considered is somewhat radical. Whenever the spaces in a table of organization are to be filled to provide a replication for any treatment, each space is filled at random from all of the available qualified personnel. In other words, there would be random sampling with replacement from a stratified finite population. It would happen, therefore, that a given subject would serve in a number of experimental military units during his participation in the experimental program. He would contribute to the effectiveness score of a replication of several, perhaps of all, the experimental treatments. He might help to determine the score of more than one replication of a given treatment. In this approach the usual techniques of analysis of variance for designs not involving more than one measurement per experimental unit would be applied if they are applicable. If this approach is legitimate, it may be the best solution, especially if we desire to use a factorial design with an appreciable number of subgroups and of replications.

The major question about this approach is the legitimacy of assuming that the error components of all the scores are independent. The reason for making this assumption is the possibility that a person's behavior is strongly influenced by the behavior of the other members of the small group in which he participates. Even if a person contributes to a number of group scores, his contribution will be made under different conditions. A man may be highly cooperative when working as a member of one rifle squad and rather uncooperative when he is put into another squad. If his behavior is not consistent under all conditions, then the fact that he helps to determine more than one group score may not necessarily force the error components of the scores to be correlated.

The argument against the assumption of independence of errors lies in the fact that under certain circumstances behavior is remarkably consistent. For example, suppose that the members of a rifle platoon are firing at targets in a situation in which the total number of hits can be recorded but the hits can not be credited to particular riflemen. Here the group score is the sum of the individual scores even though the latter are not themselves recorded. Since the number of hits made by a given person is nearly constant from time to time, and there are great individual differences in this, the group scores will be statistically dependent whenever they are partly determined by the same people.

Suppose now that a mathematical model for the group score is set up which breaks it down into components in preparation for an analysis of variance. These components have no simple relationship to the individual components mentioned earlier. Now if the treatments corresponding to two scores are different and their error components are independent, then the scores are independent. Suppose, however, that the scores are dependent because of re-use of some subjects. Then it is false that both the treatments are different and the error components are independent. But by hypothesis, the treatments are different. Therefore the error components are dependent.

In other words, although sometimes there is some reason to hope that the error components of the group scores are almost independent when the subjects are used more than once, there is often good reason to expect dependence. If this is so then I can imagine only two ways to proceed.

The first is to derive a new mathematical model which will allow re-use of personnel reassigned by stratified random sampling to form new experimental units.

The second way is to determine the relationship between levels of significance claimed by the use of conventional analysis of variance and the true levels of significance. Perhaps a Monte Carlo approach may be useful here. Finally, there may be other alternatives which I have not thought of.

The thesis of this paper may now be summarized. Experimental research on military units is faced with a serious limitation on the number of subjects available. There is some question about the adequacy of conventional

experimental designs. Your help, therefore, is solicited in two respects. First, you may be able to make suggestions about the use of already available designs. Second, if all existing designs are in some sense inadequate, you may become interested in the problem, either to work on it yourself or to encourage others to do so.

It is necessary that research be done on the organization of military units. Unless adequate experimental designs are available, however, there is danger that the experimental evidence may not be sufficient to justify conclusions drawn from the data. Your help on this problem will be, I believe, a worthwhile contribution.

PROBLEMS IN ARMY FIELD EXPERIMENTATION

Lt. Col. W. L. Clement
Military Advisor, ORO

An atmosphere of urgency and timeliness surrounds all Army testing and experimenting today. As a result we find test directives which are ambitious in scope - having several objectives - which are on a large scale, encompassing divisions and corps, and which set an extremely short time limit in which to come up with firm answers. Under these circumstances it is not surprising that sometimes the answers are not good.

I am going to talk today about some of the problems which arise in this general area of tests and experiments - a related activity - and raise some questions for later discussion.

In the first place, Army testers and experimenters are usually not statisticians or experts in experimental design. Some of the problems arise from this fact. However, even when the Army man turns to the literature on these subjects, he quickly becomes engulfed in such unfamiliar terms as "correlation," "random variability," "independent variables," "regression coefficients," and the like. And the examples he finds apply to such things as roller bearings, hogs, and corn plants. In very few places can he find literature which uses his terms and his problems - weapons, units, mobility, training, and the like - and even here a very close serarch is needed. Small wonder, then, that an appreciation of valid testing is not readily apparent.

The first problem then seems to be one of communication - to relate these agricultural and industrial techniques to military operations and problems.

Apart from the general atmosphere of urgency and the need for timely answers, Army testers operate under three general principles, pointed out by Dr. Meals of ORG in a recent paper:

1. Tests must be economical.

2. Measurements must be valid and reliable.

3. Tests must be realistic.

These three principles represent three problem areas in themselves. Number 3, achieving realism, is one of the most difficult.

So much for general problems. I will now get to some more specific matters - three in fact. One is tests of an Army combat unit; two, controllability of this unit; and three, mobility of the same unit.

First, testing (not <u>experimenting</u> with) the T/O&E (Table of Organization and Equipment) of a combat unit. As an example, let us consider a tank battalion, the problem being to test it and determine its effectiveness.

Let's see what the announced mission of this unit is, as shown in the T/O&E:

"To close with and destroy enemy forces, using fire, maneuver, and shock action in coordination with other arms."

The capabilities are also listed, some of which are:

"Attack or counterattack under hostile fire."

"Destruction of enemy armor by fire."

"High cross-country mobility", etc.

I think, as testers, we are immediately struck with the lack of any quantitative terms in description of missions and capabilities. The problem becomes how to translate these terms into measured performance in the field. Major weaknesses in current tests can be traced to the method and type of measurements taken - data collected - and to the lack of realism, as mentioned earlier.

To expand a bit on these weaknesses, most of the ratings given a unit are subjective. Umpires are used freely, and unfortunately they generally interpret rather than describe what has occurred. Here are some typical examples of items which an umpire is called on to rate in a current training test:

"Was reconnaissance adequate?"

"Did the commander employ his staff properly?"

"Were control measures adequate?"

"Disposition and control of vehicles."

"Secrecy measures."

With these items as a guide, it is certainly difficult for the umpire to be objective in rating.

Achievement of realism is another problem. Some work is currently going on in developing devices which simulate aspects of combat closely. Thus, instead of having to rely on umpire decisions, the situation is somewhat realistically portrayed on the ground. There is much work to be done in this area - how to create a combat atmosphere throughout the test.

Let's now look further at the T/O&E of this battalion. Psychological Research Associates, in their work with the rifle squad organization, listed these as the categories of factors which make up a T/O&E. To briefly run through the chart, then:

T/O&E FACTORS

Independent Variables	Controlled Variables	Dependent Variables	Field Exercises
Nr of pers Composition Equipment	Training Pers Capabilities Leadership	Controllability Fire Delivery Supply Mobility etc	Tests which bring out differences in 3 caused by varying 1

Column 1 lists the T/O&E components in which we are interested - the independent variables which the test designer is familiar with.

Column 2 lists other characteristics which will affect the test, and which must be controlled.

Column 3 shows what we are trying to measure - desirable characteristics, or dependent variables.

Column 4 is reserved for the actual problems or exercies which are set up to measure 3, and to bring out the differences.

It would seem that at present in Army tests we hold Column 1 constant, combine 2 and 3, and determine the outcome in 4. We are never really sure of what in Columns 1, 2, and 3 determined the outcome in 4.

A first order of business, before launching into extensive experimentation, is, then, to develop methods by which effectiveness of existing units can be measured more accurately than at present. Techniques, gadgets, and procedures developed in testing can be directly applied to experimental work later. And the present series of Army training tests, which units are subjected to annually, offer a ready-made framework for the tester to use.

The second specific problem has to do with an experiment to measure controllability of this battalion - listed as a dependent variable, or desirable characteristic in Column 3. In order to experiment, then, we are going to vary the independent variables in Column 1, control those in Column 2, and observe and measure controllability in Column 3 through tests which we will show in Column 4.

Now to define controllability. The commander's control duties can be divided into two major elements: 1. He plans and decides. 2. He has the unit execute the plan.

The gap between 1 and 2 is bridged by control - by the commander's communicating and supervising, and these latter are the factors to be measured. In other words, we will have a series of tests to measure communication, and another series to measure supervision.

Size, in Column 1, is the first independent variable we will consider. I propose to vary size and hold composition and equipment constant, while measuring controllability; then we will vary the other independent variables in turn.

An immediate question might well be in the interests of economy and time: should we not vary all three simultaneously? If so, in the field can we practically control these variations so that we know what has affected the outcome?

Another question: at what echelon in the chain of command will we stop - at company or platoon? Mr. Eckles, a member of our Armor Group here at ORO, has pointed out that battalion commanders actually control platoons in many cases; company commanders act as message centers in some cases, transmitting the battalion commander's orders to the platoons. This should not imply that the chain of command is violated. It does suggest, however, that battalion commander's control duties do not stop at company level. As a matter of fact, I recall a sort of rule of thumb in the Army to the effect that commanders should generally be concerned with the second echelon below their level. In other words, division commanders concern themselves with battalions, and battalion commanders concern themselves with platoons. This, then, is a point which must be settled before proceedings with our experiment.

What range of sizes do we test, and how is this determined? What are the upper and lower limits - between 10 companies and 2 companies for example? We probably can arrive at a logical, practical range of sizes by querying experienced military people.

How many battalions are needed? Can we use only command echelons, or do we need the entire unit? Must we proceed through platoon and company tests first before going to battalion level, or can useful answers be obtained by approximating performance at the lower levels? These are very practical, and economical, considerations from the Army point of view.

Now let's turn to Column 2, our controlled variables. How can these actually be taken into account and controlled? How can we arrive at meaningful results which could be applicable to the various battalions which exist today in our many armored units? What is the standard for training, discipline, and leadership, and how will our experimenter arrive at this so that he can apply his results universally?

In Columns 3 and 4 we consider test designs which measure our dependent variable and bring out differences resulting from changes in the independent variables. These performance tests should be based on critical situations which will bring out these differences. Again, military opinion is probably the best source for arriving at these critical situations.

We know that our experiment must be valid; that is, should measure what we actually trying to measure. It should be standard, so that all groups participating are graded under the same conditions. Scoring should be accurate and objective, which suggests devices of some type, together with properly instructed umpires. Our scoring indices must be carefully planned, so that they actually gauge the performance witnesses. For example, in communication, percentage of critical words heard might be an index; percentage of errors might be an index to measure performance.

So much for a brief discussion of some of the problems which arise in considering an experiment to measure controllability. In order to be certain that we trigger some response from the audience, let's look at another experiment. This time we are interested in measuring mobility of our battalion.

The aspect of mobility with which we are concerned here is vehicular operability; the unit is as mobile as the number of tanks which it keeps in operation. This implies that we must consider organization and equipment used to keep the vehicles running, as well as the vehicles themselves.

Actually, at present, tank performance is indirectly reflected in the number of mechanics needed in a unit. A broad average has been taken of tank performance, and a "vehicle equivalent" has been arrived at which by rule of thumb allocates so many mechanics for so many tanks. Actually, vechile equivalents are used in drawing up T/O&E's of all units having vehicles of any type.

We intent, therefore, to investigate this vehicle equivalent to determine in what situations it does apply and what the limiting situations are.

Again, turning to Column 1 of our table, we intend to vary <u>size</u>, here meaning number of mechanics. Some of the same questions arise as before. What range of sizes? Should we vary the other independents simultaneously? What participating troops are needed? How many battalions, if any?

Looking at Column 2, how do we take into account skills, equipment, terrain, weather, type of operation, condition and age of vehicles, at the start of our experiment? How do we relate our results to the real world of battalions spread from Europe to Korea?

In Columns 3 and 4 we should include situations which measure and discriminate between performance of vehicles, tools, and mechanics - critical situations. It would seem that a series of "canned" troubles might be built into our experiment, built up realistically from data on failure frequences.

What measurement constitutes an index of performance? Perhaps time would be the best indicator.

Having asked many questions and posed several problems, I will conclude this brief discussion. Perhaps our problems can be summed up generally in the areas of (1) Communications - understanding experimental design principles. (2) Economy. (3) Valid and reliable measurements. (4) Realism.

EVALUATION OF INTERLABORATORY TESTS WITH LIMITED CONTROLS AND DATA

W. K. Murray
Watertown Arsenal Laboratories

The following discussion concerns the problem of the proper evaluation of data received in connection with some interlaboratory determinations of oxygen in titanium alloys. The problem is complicated by the difficulty of achieving proper statistical control of the experiment when the data is obtained by voluntary cooperation of a number of laboratories, each of which differs normally, to some degree, in its methods and procedures. The difficulties which have arisen in this problem are by no means unique, but are common to most interlaboratory evaluation problems. It is felt that a solution of some of the questions arising from this specific problem would have general application.

The background of the specific problem is as follows:

Since the use of titanium has developed only recently, there have been no standard accepted methods for its chemical analysis. In order to provide generally acceptable methods, a Panel on Methods of Analysis has been set up to investigate methods for the determination of each alloying element or impurity and to recommend suitable analytical procedures. In the case of most elements, procedures have been developed, tested by a number of cooperating laboratories and found to be quite satisfactory with regard to precision and accuracy.

In the determination of oxygen in titanium, however, no procedure has yet been adopted and recommended for general use. One reason for this is that there are no standard specimens available containing known amounts of oxygen against which procedures can be tested.

As a preliminary investigation, it was decided to limit our analysis to two general sources of variation: that due to the samples and that due to the laboratories. It is believed that, if we can show interlaboratory differences to be the significant source of variation, our problem would be reduced to a study of laboratory methods.

The samples consisted of commercial titanium and titanium alloys available in stock, thus eliminating any control over their preparation. The cutting of the original material and randomizing of the samples for distribution to the different laboratories is the first control we are able to exercise over the samples in this design. The samples were distributed to the cooperating laboratories, who were requested to make four determinations for each titanium alloy using one or both of two suggested methods; the number of determinations were restricted due to the cost involved. Homogeneity of the sample being unknown, we attempted by randomization to reduce the influence of oxygen segregation in the samples.

The hypotheses we wish to test are: (1) there is no within-sample variation; (2) there is no between-laboratory variation; and (3) the two methods tested give similar results. The purpose of this study is to determine whether the difference in results is due to differences among laboratories or to segregation in the titanium samples; and, if possible, to determine whether a technique for determining oxygen in titanium is suitable for recommendation as an acceptable procedure.

After this general statement of the problem, we should like to mention some of the specific questions which have arisen and which must be resolved if a logical statistical approach is to be utilized.

Preliminary to any statistical analysis one must handle the question of rejecting data. In an experiment such as this one, which is to some degree uncontrolled, this is an important point. Certain laboratories are personally known to be more reliable than others by virtue of better equipment, more experience and other factors. Can one give more weight to the results of these laboratories than the others and still avoid biasing the results by personal prejudices? In our case, it is very tempting to eliminate the results of about half of the thirteen cooperating laboratories. Previous experience has indicated that there is a group of laboratories whose work is more reliable than the others. These laboratories, in this testing program, agreed with each other much more closely than did the other laboratories. Yet, on purely statistical grounds, there is no reason to eliminate more than one laboratory on the basis of the results received.

Another question concerns the analysis of data gathered employing two different analytical procedures in the same laboratory or in different laboratories. Should the methods be compared on a laboratory to laboratory basis or should the results be combined by method? Also, under what conditions can the laboratory results from different specimens be combined to investigate differences between laboratories and between methods?

There are specific difficulties, but we believe a general discussion of the attitudes and aims that one should have when confronted with a problem such as this in which the controls and data are limited would be appropriate. What, for instance, should be the major concern of a statistical treatment which is the first attempt to exercise statistical control on the variables under consideration?

DESIGN OF EXPERIMENT

A. Bulfinch
Picatinny Arsenal

Engineers and scientists who have recently been introduced to the subject of statistics, often ask: "Just what does one do to design an experiment in the modern statistical sense?" This is a good question, and there should be a sensible answer that the engineer can understand and use. An examination of the literature shows that much has been written on the subject, but no unified procedure that can be identified as such can be found in any one document. Too many books have been written for statisticians and too many handbooks contain only methods of analysis.

The engineer would like something tangible to manipulate, or a set of instructions that can be followed, something short of book length. The statistician may say that this is impossible! But his conclusion is based on the assumption that the engineer is completely ignorant of the subject of statistics, and that to use statistics one must know all of the designs and techniques. Experience has shown that this is not true. Many engineers and scientists will design the most efficient experiment by using just good common sense. Any one job requires the use of only a few techniques, not the whole spectrum. From this I have concluded that an explicitly described, unified design-of-experiment procedure would be useful to engineers. Such a description may include terms not familiar to the engineer or scientist, but an effort to understand the definitions of these terms would be the shortest route to a working knowledge of the design of experiment in the modern statistical sense.

Planning an experiment along statistical lines forces one to consider what it is he is seeking and what steps are required to obtain it. This often leads to the recognition of pitfalls and fallacies in advance of data collecting.

The "design of experiment" is essentially the pattern of taking observations. In its broader sense this procedure also includes the analysis of results. The object of designing an experiment in the modern statistical sense is two fold.

1. To obtain economy of experimentation. That is, to insure that essential information is obtained with minimum cost in time and effort. "Essential information" is defined as information such that additional data will not change the conclusions drawn, in a practical sense.

2. To obtain a "yardstick" with which to evaluate the results. This "yardstick" is called the experimental error, which is obtained by replicating the results.

The "design of experiment" may be regarded as an aspect of the scientific method. The intrinsic characteristics of the scientific method are the examination of what is known and the formulation of theories or hypotheses which may be verified by experimentation. The concept of experimentation is the crux of the entire matter, for any question whose answer may not be obtained by planned observations is not in the realm of science.

The actual formulation of hypotheses and theories is a matter of intuition, native ability, and insight. Verification of these hypotheses and theories cannot be absolute, for we can only show that the observations are compatible with the hypothesis within the limits of experimental error. This is the major reason for the use of the "null" hypothesis in statistics. We make changes and assume or theorize that these changes have made no difference, that the difference is "null" or amounts to nothing. In every case we state our questions to be answered by the experiment in a hypothesis to be disproven by the data. If we fail to disprove the hypothesis, then we accept it as true or reserve decision. This means we have three alternatives: reject the hypothesis, accept the hypothesis, or reserve decision. In the analysis of variance (of designed experiments) we combine the last two alternatives and state: "There is not sufficient data to detect a difference".

The hypothesis that there is no difference (the null hypothesis) is unrealistic, since different treatments must have produced some difference. The real problem is to obtain estimates of the magnitude of the difference and determine whether this has any practical or economic importance.

The acceptance of any hypothesis on the basis of data obtained from samples of a population or universe is subject to a probability of error. This principle represents the basis of modern statistical theory. In testing a hypothesis there are two possible errors: Type I Error is the risk of rejecting the hypothesis when it is true. Type II Error is the risk of accepting the hypothesis when it is false. The value of designed experiments is that they minimize these risks of error with minimum effort. That is, statistically designed experiments are the most efficient experiments since they can obtain essential information with minimum cost.

A hypothesis must provide the answer for a practical problem, provide an explanation of known facts, and give predictions that can be verified. It is essential that hypotheses and their outcomes be formulated before verification is attempted. Valid probability statements cannot be made about statistical tests suggested by the data to which they apply.

The theory of statistics, which is entirely deductive, provides a basis for inductive processes. No inductive inference is certain to be correct, so every conclusion drawn from finite experimental data is subject to error. With the aid of mathematical statistics, probability statements may be made about these errors.

The role of statistics in the scientific method has three functions:

1. Description - This is the reduction of a mass of data to such quantities as the mean and the variance. If the data is all of the relevant information about the whole population, these quantities are called parameters and the description is deductive. If the data is only a sample of the whole population, these quantities are called statistics and the decription is inductive.

2. Analysis - This means, given observed values from a sample, to

estimate the population parameters. Also analysis can mean given observed values from two samples, to determine whether the two samples came from the same population.

3. Prediction - This means rational inductive processes. This is the major objective of the application of the scientific method to natural phenomena. The practical application of the theory of probability through the use of statistical techniques has made it possible to make predictions from controlled experiments with mathematical precision.

Emphasis should be placed on the application of the theory of probability since at the theory level academic sterility is an ever present danger. As Bross puts it, "Academitis is a disease characterized by hair-splitting and eventually, rigor mortis."

For our purposes it is useful to distinguish between two types of experiments.

1. The determination of the numerical magnitude of a particular characteristic for a specified population.

2. The determination of the effect of two or more treatments on a particular population characteristic.

In the first type the populations consist of existing items or properties, and it is simply a matter of measuring them. In the second type the populations studied are created by the experimenter in the act of taking measurements. It is in this latter type of experiment that statistical design techniques are required.

Planning the experiment in advance of data collecting cannot be overemphasized. In the past, an experiment was considered a venture into the unknown, and as such, any approach and any result was acceptable, since neither could be predicted or evaluated. This was a boon to the experimenter and gave him a free hand. But modern techniques have changed all this by furnishing systematic procedures for designing experiments and analyzing the results. Inefficient methods and unreliable data can no longer be tolerated.

Described below are some of the things that should be done in planning an efficient experiment and analyzing the results. This is what I believe engineers want when they ask, "How can I design an experiment?" and what the literature has glossed over:

a. Plan your experiments well. The conclusions and inferences that can be drawn depend on the way in which observations are made.

b. Use common sense. Don't accept results which contradict common sense.

c. Use all available knowledge and information from past experience.

d. Consider all possible sources of error. List the variables to be controlled, those to be varied, and the levels of those to be varied.

e. Consider the entire scope of the problem. Without regard to cost, time, or effort, consider what it is you would like to know eventually. If this turns out to be a very large experiment, consisting of many variables, or a very expensive experiment the cost of which is prohibitive, divide the whole problem into rational parts. This makes possible a systematically-planned approach. It also makes it possible to relate your statistical design to cost and the amount of information required.

f. Consider all possible outcomes, and their physical interpretation. Results that have no physical interpretation have no practical value.

g. Choose carefully the criterion on which conclusions will be based. Density results are of little value if the use of the material depends upon the melting point.

h. Randomize sample specimens. This can be done by using tables of random numbers or by drawing numbers out of a hat. In any case, randomization insures better representative samples and guards against biased results.

i. A valid estimate of experimental error must be obtained with which to evaluate the results. This can usually be done by taking repeated measurements under the same controlled conditions. This is called "replication".

j. The sample size (the number of repeated measurements under the same controlled conditions) should be adjusted to control the alpha and beta errors. The alpha error is the risk of rejecting good material, the Type I error, or the producer's risk. The beta error is the risk of accepting poor material, the Type II error, or the consumer's risk. In order to control these errors, some knowledge of the variability (experimental error) must be available. In addition, a decision must be made concerning the magnitude of the difference that must be detected to make the experiment economically feasible.

k. Carefully formulate the questions to be answered. Develop the right hypotheses by asking the right questions which the experimental results are expected to answer. To show conclusively that process A gives a higher yield than process B, is of little value if neither produces a usable product.

l. Of the many experimental designs available, choose the one that fits your particular problem requirements. Factorial designs are very efficient since they will provide complete information about all of the variables, as well as their interrelationships, with only a fraction of the work required by the classical one-at-a-time procedure. This type of design is particularly useful when little is known about the system being studied, or when it is known that there is a very complex relationship

among the variables. If the number of variables to be studied exceeds 5 or 6, designs such as the Latin square and fractional factorials should be considered to affect further economies of experimentation. These latter designs are also useful for a sequential approach to a problem containing more than 5 or 6 variables of interest. The analysis of regression, the analysis of covariance, and the method of confounding, are useful when there are variables that cannot be controlled. The correlation coefficient and the analysis of regression are useful in studying the relation between variables -- such as cause and effect.

m. A property of these designs, known as Orthogonality, should be controlled in order to simplify the calculations and the interpretation of the results. This property insures that all the variables (called main effects) and all of their interrelationships (called interactions) can be independently estimated without entanglement.

n. Care should be taken so that the effect of one variable is not confounded or confused with that of another when independent measurements of each are required. Little can be concluded about the moisture content of two products, made by different processes, if ambient humidity conditions are permitted to effect the results. In such a case, the moisture content due to the process is confounded (or confused) with that due to the humidity. If the ambient humidity condition is an important variable in the system, it should be controlled and the experiment designed to determine its effect. If it cannot be controlled, the experiment should be designed so that changes in humidity can affect only unimportant parts of the experiment, such as the higher order interactions.

o. The concept of interaction should be understood. Interaction is said to be present when certain particular combinations of conditions produce unusual results. This is the nonadditive or unpredictable portion of the experiment, and, as such, is the only patentable portion of the experiment. There can be interaction between two or more factors (variables). Interactions involving three or more factors are referred to as the higher order interactions. Interactions involving five or more factors seldom have any physical interpretation or practical importance.

p. The observations or measurements must be independent for many designs. Measurements are said to be independent if the probability that one of them will have a certain value is the same, no matter what values are obtained for other measurements. This means that the results cannot be correlated and that the taking of a measurement will not affect the outcome of succeeding measurements. For example, if the first measurement raises the temperature of the system, and the results are affected by temperature changes, then the probability of reproducing the first result with a second measurement is nil. In such a case, the temperature must be controlled in order to obtain independent measurements. However, if the variables are correlated, the analysis of regression or covariance can be used.

q. There must be assurance that the error of measurement (called

the variance) does not change from one portion of the experiment to another. That is, we must comply with the requirement of homogeneity of variances. This is important because there are two sources of variation -- the means (or averages) and the variances. If we observe a difference, we want to be in a position to determine whether it is due to the means or variances. We are usually interested in changes of the mean values, so if the variances are constant or homogeneous and we observe a change, we will be able to conclude that it is due to the means.

r. The concept of degrees of freedom should be understood, since it is used extensively in the analysis of data. The number of degrees of freedom is equal to the number of independent observations minus the number of parameters (such as the means) estimated. In computing the variance, for example, only (n-1) of the deviations from the mean can be independent. The nth deviation has to be restricted in order to make all "n" deviations add up to zero.

s. The type of measurement to be used should be considered for the sake of efficiency. Variable type data is data that can vary from minus infinity to plus infinity on a continuous scale. This type of data furnishes the most information per observation. Attribute data is qualitative type data and consists of discrete entities. Attribute data is sometimes called "go" "no go" data. The latter kind of data gives the least information per observation.

t. The assumption of normality must be considered, since most probability statements are based on this assumption. However, if you are dealing with the distribution of averages or with small sample sizes, the question of normality is purely academic for the following reasons:

(1) The distribution of all averages can be considered normal, regardless of the source of the individual values -- especially averages of four or more values.

(2) No reliable test of normality is available for small sample sizes. In addition, there are robust tests now available which are insensitive to deviations from normality.

The numerical values of measurable properties of products manufactured under controlled conditions can be considered normally distributed. The F-test in the analysis of variance, and the t-test for the difference between two averages are both insensitive to deviations from normality. With this in mind, it can be concluded that the assumption of normality is sufficiently valid for most practical purposes, unless there is definite information to the contrary. At worst, your level of probability will be low by a few percent.

u. The saving of time and effort through the use of statistically designed experiments can be demonstrated by the following comparison with the classical one-at-a-time procedure.

Classical Procedure

	$Temp_1$	$Temp_2$
$Press_1$	-	-
$Press_2$	-	-

Statistical Procedure

	$Temp_1$	$Temp_2$	Averages to compare effect of pressure
$Press_1$	-	-	-
$Press_2$	-	-	-
Averages to compare effect of temp	-	-	

In the above illustration let a dash mark represent a single determination.

Classical Procedure:

The effect of temperature is determined by comparing the average of duplicate determinations at each of the two temperatures for the first pressure level. We repeat the process for the second pressure level. To determine the effect of pressure we compare the average of duplicate determinations at each of the two pressures for the first temperature level and repeat the process for the second temperature.

Statistical Procedure:

The effect of temperature is determined by averaging over the two pressure levels. That is, the value obtained for the condition of "temperature one" and "pressure one" is averaged with the value obtained for the condition of "temperature one" and "pressure two". The process is repeated for "temperature two". The two averages obtained in this way are compared to determine the effect of temperature. The effect of pressure is determined in a similar way by averaging over the two temperature levels.

In both cases we were comparing averages of duplicate determinations, but in the statistical procedure we attained this precision with only half the number of determinations used in the classical procedure. This economy is made possible by removing two long-standing barriers, namely:

1. You can't average "apples and pears".

2. You can't vary more than one thing at a time.

The removal of these barriers and using each measurement or determination for more than one purpose is mathematically possible if we assume that the "error" created by changing the pressure in taking a measurement at "temperature one" is equal to the "error" created by changing the pressure in taking a measurement at "temperature two". If the effect of these two factors upon each other is additive, this assumption is valid. By additive is meant that if changing the pressure a given amount produces

a 15% increase in yield at "temperature one", changing the pressure the same amount at "temperature two" will also produce a 15% increase in yield.

Algebraically, if:

$$A = B,$$

$$(A + C) = (B + C),$$

$$A - B = 0$$

$$(A + C) - (B + C) = 0$$

This means that the error due to changing the pressure when measuring the effect of temperature will cancel out, since measuring the effect of a factor (or variable) is actually a process of subtraction and an evaluation of the difference.

If there are interaction (nonadditive) effects present, the above additive relation still holds, but additional work must be done to separate them from experimental error. This can only be done with statistical procedures. Interaction can never be measured or calculated with the classical procedure.

One of the major objectives of the statistical procedure is to obtain a measure of experimental error (or reproducibility) with which to evaluate the main factor and interaction effects so that variation due to chance alone can be distinguished from differences due to assignable causes.

To get a measure of experimental error, at least duplicate determinations must be made for each condition. In the above example this would require doubling the number of determinations in the experiment under "Statistical Procedure". This would now mean that we could compare averages of four determinations. To make the experiment under "Classical Procedure" comparable, we would have to double the number of determinations here also in order to compare averages of four determinations.

Now a detailed comparison of the two procedures shows a wide divergence in favor of the "Statistical Procedure". By means of this procedure the total error in the above two-factor experiment can be divided into five components:

1. Main effects.
   a. Temperature.
   b. Pressure.
2. Interaction

3. Experimental error.

   a. Replication.

   b. Residual error.

It is assumed that the residual error is that portion of the totel error which remains after all the error due to assignable causes has been removed. That is, the residual error is assumed to be due to chance causes alone. As such, the residual error is used as a yardstick to evaluate the main and interaction effects through the use of the F-test. This test is a mathematically precise method for evaluating data to distinguish between variations due to chance alone and differences due to assignable causes.

In contrast, the "Classical Procedure" includes no means of determining:

1. Most efficient and economic experimental designs.

2. Interaction effects.

3. Residual error.

4. Difference between variations due to chance alone and differences due to assignable causes.

The result of these deficiences leaves only common sense and subjective judgment (with all the attendant personal biases) to design experiments and analyze data in the "Classical Procedure".

To demonstrate more clearly that more than one thing at a time can be varied in the "Statistical Procedure", the following fractional factorial design is presented:

	A_1		A_2	
	B_1	B_2	B_1	B_2
C_1		-	-	
C_2	-			-

Measurements are made for only those conditions indicated by the dashes; yet the effect of all three factors can be determined and evaluated if there are no significant interactions present. This is only one-fourth the amount of work required to obtain the same precision by the "Classical Procedure". Truly a saving of time!

LINEAR MODELS IN THE ANALYSIS OF VARIANCE*

M. B. Wilk
Princeton University

Introduction. In recent years a new word has won widespread acceptance into the technical language of statistics. I have in mind the term "robust". This expression was introduced by Box [1] to characterize statistical tests which are not overly sensitive in their behavior and meaning to preliminary statistical assumptions. What he meant us to understand by this word is strongly suggested by its dictionary definition (Webster's 2nd Edition):

> "having or evincing strength or vigorous health; strong; muscular; vigorous; sound."

While the use of the word in statistics is new, the basic concern which it reflects is not at all recent. For example, the introduction by Fisher [3] of the device of deliberate randomization in experimentation was motivated by a desire to provide a robust basis for statistical inference. Similarly, for many years so-called non-parametric or distribution-free procedures have been advocated to relieve inferences from the weight of assumptions whose justification may be difficult or impossible.

In addition to our explicit concern with the relative robustness of significance tests and estimation procedures, I would like to direct some attention to the question of robustness of statistical experimental designs and of statistical models.

As a simple example of non-robust experimental procedure consider the situation suggested by (1).

$$y = f(x;\ \alpha,\ \beta,\ \gamma,\ \ldots) + e. \tag{1}$$

If it is known that the functional relation is given by (2),

$$y = \alpha + \beta x + e, \tag{2}$$

then we know that, with moderately reasonable statistical properties of the errors, a "best" selection of values x_i at which the responses y_i should be observed would be such as to maximize (3), which measures the dispersion of the x_i values.

$$\sum_i (x_i - \bar{x})^2 . \tag{3}$$

---

*A talk given at the Second Conference on the Design of Experiments in Army Research Development and Testing, Washington, D. C., October 19, 1956. Prepared in connection with research sponsored by the Office of Ordnance Research.

This effectively means that the preselected x_i values should be concentrated at the two extremes of the possible range of x. Clearly this design is not at all robust since if there is, in fact, some curvature in the relation between y and x, as for example in (4)

$$y = \alpha + \beta x + \gamma x^2 + e, \tag{4}$$

then we could get no clue of this from an experiment with all x_i values at the two ends.

As another example, consider the relationship of randomized complete blocks and incomplete blocks designs. In the latter designs the presence of unanticipated interactions cannot, in general, be easily detected and may in consequence introduce serious errors into conclusions. In this sense, complete blocks are more robust than incomplete blocks. On the other hand, the use of complete blocks may lead to overly large uncontrolled variation, with consequent concealment of effects of interest. Similarly, fractional factorial designs will, in general, be less robust than full factorial designs in that the confounding which occurs in the fractionated designs may be of importance and go undetected.

In contrast, one of the arguments given by Fisher [4, p. 106] in support of factorial experiments is as follows:

> "Any conclusion has a wider inductive basis when inferred from an experiment in which the quantities of other ingredients have been varied, than it would have from any amount of experimentation, in which these had been kept strictly constant."

The remainder of this paper is devoted to classification models and regression models, with particular reference to their robustness characteristics. My intention is to try to deal with general ideas and principles rather than to attempt to convey any detailed methodology.

Analysis of Variance or Classification Models. I am sure everyone here is familiar with models of the general appearance of (5).

$$y = \mu + a + b + c + \ldots + e. \tag{5}$$

Such models have been used increasingly widely in the past decade as a basis for justifying the analysis of variance. It so happens that if one makes some suitably chosen assumptions concerning this model, it is possible to provide an elegant and rather complete mathematical-statistical justification for the analysis of variance. Unfortunately, this justification does not require any deep-rooted scrutiny of the meaning or possible origin of the model. Due perhaps to the abstract treatment of these models, there have occurred some conflicting views on appropriate interpretation of fairly simple experimental situations, such as the mixed model case of a two-factor experiment. The heart of this controversy lay in the treatment of the same experimental situation in terms of different, arbitrary assumptions concerning the components of the model.

It seems apparent that if these linear analysis of variance models are to be useful for a wide range of experimental circumstances, then they must have a robust status in the sense that they must derive their meaning and properties not from arbitrary assumptions but rather from a very general framework or concept of experimentation as a means of learning about the real world, combined with such direct properties as the experimental design itself possesses. The model must not depend on very special properties of specific experimental situations.

Consider the essential ingredients of a simple two-factor experimental situation. In such a situation, idealized, one would be concerned with determining the effects on some response Y which are attributable to variation in the levels of each of two factors, $\mathcal{A}$ and $\mathcal{B}$. Clearly, this description is grossly incomplete, even for an idealized framework, for no provision has been made for the implicit background or surroundings. To account for some of this we introduce the notion of experimental units. For example, a chemical engineer might wish to study the effect of column diameter and type of packing on the maximum throughput in a packed column. Here the response is to be maximum throughput, perhaps in pounds per hour or more likely in pounds per square foot per hour; the factors (or independent variables) are column diameter and packing; and the experimental units will summarize such features as the method of determining the maximum throughput, the changes which occur in the fluids and equipment employed, uncontrolled ambient temperature and pressure changes, and so on. Clearly some properties of the experimental units will be, essentially, constant for all units, while other characteristics will fluctuate from one to the other.

Suppose factor $\mathcal{A}$ to have A levels and factor $\mathcal{B}$ to have B levels, and let the indices i and j have range as given in (6).

(6) $$i = 1, 2, \ldots, A$$

$$j = 1, 2, \ldots, B.$$

For initial simplicity let us assume that all experimental units are identical. Then it is reasonable, in many cases, to conceive of a number Y_{ij}, defined in (7), namely

(7) Y_{ij} = true or typical response which would be observed from the treatment combination consisting of the ith level of factor $\mathcal{A}$ and the jth level of factor $\mathcal{B}$.

If we now use dots to denote means or averages, as exemplified in (8)

(8) $$Y_{i.} = \frac{1}{B} \sum_{j} Y_{ij} ,$$

then we can write the algebraic identity given in (9).

(9) $$Y_{ij} = Y.. + (Y_{i.} - Y..) + (Y_{.j} - Y..) + (Y_{ij} - Y_{i.} - Y_{.j} + Y..)$$

$$= \mu + a_i + b_j + (ab)_{ij} .$$

It is apparent from their definition that the components of this population model can be given a physical interpretation or meaning. This meaning is suggested by the nomenclature defined in (10).

(10)
μ is the overall mean,

a_i is the main effect of level i of factor $\mathcal{A}$,

b_j is the main effect of level j of factor $\mathcal{B}$,

$(ab)_{ij}$ is the interaction of level i of factor $\mathcal{A}$ with level j of factor $\mathcal{B}$.

Two important aspects of these defined components of the population model should be made explicit. First, the definition of, for example, the main effects of factor $\mathcal{A}$ depends crucially on which levels of factor $\mathcal{B}$ are included in the experimental situation. Second, the relative and absolute magnitudes of the interactions will depend on the scale of measurement of the responses Y. Thus the same two factors may show important interaction on one scale of response Y, and yet may show negligible interaction on some other scale of response; for example, $g(Y) = \sqrt{Y}$. For the very special and important case in which interactions are negligible then the meaning of the main effects of factor $\mathcal{A}$ become independent, in general, of the levels of factor $\mathcal{B}$ involved. This is formally stated in (11), which follows directly from the definition of $(ab)_{ij}$.

(11) All $(ab)_{ij} = 0$ implies $Y_{ij} - Y_{.j} = Y_{i.} - Y_{..} = a_i$.

It is, however, worth repeating that the relative size and importance of the two-factor interactions depends not only on the mechanics of the situation but also on the scale in which the responses are analyzed.

The same notions may be extended to the more realistic case where experimental units are different; that is, where unperceived or uncontrolled variation in the background may condition or obscure our evaluation of the effects of the factors. The population model then takes the form given in (12).

(12) $$Y_{ijk} = \mu + a_i + b_j + (ab)_{ij} + e_k + p_{ijk}.$$

In this expression, e_k may be called the additive unit error and p_{ijk} the interactive unit error. The population model components are now defined with respect to the relevant population of experimental units and of treatment combinations. The e_k reflect variation among experimental units, averaged over all treatments. The p_{ijk} reflect interactions of treatment combinations with experimental units.

Now as yet we have said nothing about an actual experiment; we have simply developed a formal framework which we hope is sufficiently flexible to fit most two-factor experimental situations reasonably well.

Suppose a factorial experiment is now carried out, as sketchily outlined in (13).

(13)
(i) Select a levels of factor $\mathcal{A}$; $a \leqq A$.
(ii) Select b levels of factor $\mathcal{B}$; $b \leqq B$.
(iii) Have r replications of the selected a x b treatment combinations.

At this point it is necessary to inquire just how selection of levels and allocation of experimental units is to be made. To the extent that physical randomization (i.e., random numbers) is employed, objective statistical-probability ideas can be used to make inferences from the actual experimental observations to certain fairly well defined broader populations. To the extent that randomization is not employed, broader inferences can not be based solely on statistical-probability notions.

If we have conformed "to all the principles of allowed witch-craft" -- to use a phrase due to W. S. Gosset, better known as 'Student' -- we can carry our population models forward to a statistical model for the observations. Use the notation defined in (14).

Let

$$u = 1, 2, \ldots, a$$

$$v = 1, 2, \ldots, b$$

denote selected levels of factors $\mathcal{A}$ and $\mathcal{B}$, in order of their random selection;

(14)
$$f_{uv} = 1, 2, \ldots, r$$

denote replication of treatment (u, v);

x_{uvf} represent the observation from replication

$f = f_{uv}$ of treatment (u, v).

Then we can write a statistical model for the observations x_{uvf} in the form given in (15).

(15)
$$x_{uvf} = u + \alpha_u + \beta_v + (\alpha\beta)_{uv} + \varepsilon_{uvf} .$$

This model derives from the population model by imposing the conditions of the experimental design, including the randomization employed as well as the pattern. An outline of the relationship is given in (16) using the simplifying assumption that all p_{ijk} are negligible.

Define the following design random variables:

$\alpha_i^u = 1$ if selection u corresponds to i in the population of levels of $\mathcal{A}$,

$= 0$ otherwise.

$\beta_j^v = 1$ if $v \longleftrightarrow j$,

$= 0$ otherwise.

(16) $\delta_k^{uvf} = 1$ if the fth replicate of selected treatment (uv) falls on experimental unit k,

$= 0$ otherwise.

The properties of these random variables derive from the pattern of random selection and allocation (i.e., the experimental design) employed.

We then have, with the simplifying assumption that $p_{ijk} = 0$,

$$\alpha_u = \sum_i \alpha_i^u a_i \; ; \quad \beta_v = \sum_j \beta_j^v b_j \; ;$$

$$(\alpha\beta)_{uv} = \sum_{ij} \alpha_i^u \beta_j^v (ab)_{ij} ; \quad \varepsilon_{uvf} = \sum_k \delta_k^{uvf} e_k \; .$$

The important point is that the properties of the components of the statistical model for the observations follows from combination of the population model (which was based on the rather general concept of a true response) with the experimental design which is actually imposed by the experimenter.

The implications of this model so far as interpretation of the analysis of variance is concerned is partially indicated by the expectations of mean squares given in Table 1.

Table 1

Due to	d.f.	M.S.	E.M.S.
$\mathcal{A}$	(d-1)	A*	$\sigma_e^2 + r\frac{(B-b)}{B}\sigma_{ab}^2 + rb\,\sigma_a^2$
$\mathcal{B}$	(b-1)	B*	$\sigma_e^2 + r\frac{(A-a)}{A}\sigma_{ab}^2 + ra\,\sigma_b^2$
$\mathcal{A} \times \mathcal{B}$	(a-1)(b-1)	I*	$\sigma_e^2 + r\,\sigma_{ab}^2$
Residual	ab(r-1)	R*	σ_e^2

Definitions: $\sigma_a^2 = \frac{1}{A-1}\sum_i a_i^2 \; ; \quad \sigma_b^2 = \frac{1}{B-1}\sum_j b_j^2 \; ;$

$$\sigma_{ab}^2 = \frac{1}{(A-1)(B-1)}\sum_{ij}(ab)_{ij}^2 \; ; \quad \sigma_e^2 = \frac{1}{P-1}\sum_k e_k^2 \; .$$

It can be seen from Table 1 that if B = b, that is, if all levels of factor $\mathcal{B}$ in the population considered are studied in the experiment, then the component of variation due to interactions, σ_{ab}^2, does not contribute to the $\mathcal{A}$ mean square. Contrariwise, if B>> b* so that only a small proportion of possible levels of factor $\mathcal{B}$ are sampled and one wishes to make inferences relative to the entire population of levels of factor $\mathcal{B}$, then the interaction component of variation does contribute, on the average, to the $\mathcal{A}$ mean square.

The fact that the results of Table 1 derive from the quite robust model we have developed is one strong indication that the analysis of variance is a meaningful procedure, without regard to more sophisticated assumptions.

The results given in Table 1 involved the simplifying assumption that the interactive unit errors, the p_{ijk} which measure unit-treatment interactions, were negligible. Moreover, the model used contained no provision for either measurement errors or variabilities in preparation of treatments. The results on expectations of mean squares under a more general model, which do provide for such effects, are given in Table 2, with a notation that lends itself readily to extension to more complex situations.

Table 2

Due to	E.M.S.
$\mathcal{A}$	$\Sigma_o + r\Sigma_{ab} + rb\Sigma_a$
$\mathcal{B}$	$\Sigma_o + r\Sigma_{ab} + ra\Sigma_b$
$\mathcal{A}\times\mathcal{B}$	$\Sigma_o + r\Sigma_{ab}$
Residual	Σ_o

Definitions:

$$\Sigma_a = \sigma_a^2 - \frac{1}{B}\sigma_{ab}^2 - \frac{1}{P}\sigma_{ae}^2 + \frac{1}{BP}\sigma_{abe}^2$$

$$\Sigma_b = \sigma_b^2 - \frac{1}{A}\sigma_{ab}^2 - \frac{1}{P}\sigma_{be}^2 + \frac{1}{AP}\sigma_{abe}^2$$

$$\Sigma_{ab} = \sigma_{ab}^2 - \frac{1}{P}\sigma_{abe}^2$$

$$\Sigma_e = \sigma_e^2 - \frac{1}{A}\sigma_{ae}^2 - \frac{1}{B}\sigma_{be}^2 + \frac{1}{AB}\sigma_{abe}^2$$

$$\Sigma_{ae} = \sigma_{ae}^2 - \frac{1}{B}\Sigma_{abe}$$

$$\Sigma_{abe} = \sigma_{abe}^2 - \frac{1}{(A-1)(B-1)(P-1)}\sum_{ijk}(p_{ijk}-p_{i.k}-p_{.jk})^2$$

$$\Sigma_o = \sigma^2 + \Sigma_{abe} + \Sigma_{ae} + \Sigma_{be} + \Sigma_e$$

σ^2 = Variance of "technical errors"

P = size of population of experimental units.

* The symbol >> is used to denote "much larger than."

A close inspection of the results of Table 2 will show that the existence of unit-treatment interactions occasions a bias in the analysis of variance in the sense that unbiased estimates of error cannot in general be obtained. If the size of the relevant population of experimental units is large, however, this bias is negligible.

Thus we see that, with appropriate interpretation, classification models can be given a robust status. Such models can be used whenever factor levels are distinguishable either qualitatively or quantitatively. They help in several ways: (1) They provide a formal structure whose relation to the populations of interest is usually well-defined. This helps in interpreting the actual experimental results in terms of the broader populations of concern. (2) Properly used and interpreted, these models help provide insight into the physical meaning of terms such as "effects" and "interactions of factors." (3) The use of the models brings out into the open the necessary assumptions (or conditions) which may be necessary for an unambiguous interpretation of the analysis of variance. In the same way they help in evaluating the possible direction of misinterpretation if assumptions fail. (4) By appropriate statistical analysis — as, for example, by finding a scale for analysis on which interactions are negligible — we may be led to simplified and hence more developed models.

The main deficiency of these general classification models -- and it is overwhelmingly important -- is that the classification models do not directly concern themselves with functional relations between response and factors or independent variables. If quantitative information on factors is available, the use of a classification model will simply ignore this information -- obviously an undesirable feature. Thus, as ordinarily employed, classification models when properly interpreted do not require sophisticated information to be useful, but by the same token they do not lead to sophisticated insights.

Further published work on classification models can be found in references [2], [5], [6], [7].

Polynomial Regression Models. Another type of model is widely used in statistical analysis of experimental data is the polynomial regression model, such as in (17).

$$y = \alpha_{00} + \alpha_{10}z_1 + \alpha_{02}z_2 + \alpha_{11}z_1^2 + \alpha_{22}z_2^2 + \alpha_{12}z_1z_2 + \text{error}. \tag{17}$$

It is easily seen that such models, as well as the classification models, can be put in the form of a linear multiple regression model such as (18).

$$y = \beta_0x_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_kx_k + e. \tag{18}$$

The appropriate correspondence for regression models is indicated in (19).

$$\begin{aligned} &\alpha_0 = \beta_0 \;;\; 1 = x_0 \,, \\ &\alpha_{10} = \beta_1 \;;\; z_1 = x_1 \,, \\ &\alpha_{02} = \beta_2 \;;\; z_2 = x_2 \,, \\ &\alpha_{11} = \beta_3 \;;\; z_1^2 = x_3 \,, \\ &\text{etc.} \end{aligned} \tag{19}$$

To show the formal connection for classification models, consider for example a 2×2 factorial experiment. We could write a simplified classification model as in (20).

$$
\begin{aligned}
y_{11} &= \mu.1 + a_1.1 + a_2.0 + b_1.1 + b_2.0 + e_{11} , \\
y_{12} &= \mu.1 + a_1.1 + a_2.0 + b_1.0 + b_2.1 + e_{12} , \\
y_{21} &= \mu.1 + a_1.0 + a_2.1 + b_1.1 + b_2.0 + e_{21} , \\
y_{22} &= \mu.1 + a_1.0 + a_2.1 + b_1.0 + b_2.1 + e_{22} .
\end{aligned} \tag{20}
$$

Clearly this has the same formal structure as the multiple regression model, with the x's taking on the values 0 and 1, appropriately, and the parameters of the classification model playing the role of regression coefficients.

While this formal identification is sometimes convenient in allowing a certain unity and elegance in mathematical developments concerning least squares and analysis of variance theory, there are important logical and practical distinctions between classification and regression models.

In a regression model such as (17), the values of z_1 and z_2 are quantitative identifications or descriptions of the levels of two factors under study, and it is ordinarily implicit that the values of the z's are sufficient to summarize the important characteristics of the actual factor levels used in the experiment, in the sense, for example, that we ordinarily believe the application of a particular pressure to be summarized by the number of pounds per square inch associated with the applied pressure.

While it is basic in a regression model that the factors be quantified, their quantification known, and that the numerical measure be a complete summary, the classification model does not require this information. On the other hand, even for comparatively simple experimental situations the number of parameters in a classification model can rapidly become very large indeed. For example, if in a $5 \times 5 \times 5$ factorial experiment we could ignore three-factor interactions but no others, we would need 61 independent parameters in a classification model. A moderately complex regression model might employ 20 parameters. Clearly the classification model assumes less, but also accomplishes less.

It has been said of the popularity of the assumption of normal or Gaussian distribution that "everybody believes in the law of errors, the experimenters because they think it is a mathematical theorem, the mathematicians because they think it is an experimental fact." One wonders whether a similar remark might not be appropriate to the popularity of polynomial regression.

The basic mathematical theorems are due to Taylor and to Weierstrass. Taylor's Theorem tells us that if a function $f(x)$ has derivatives of order k, then $f(x)$ may be expanded as a power series of the form shown

in expression (21). In this expression x_0 is some preselected value of x, and R_n is the remainder after n terms of the expansion. The coefficients $f'(x_0)$, $f''(x_0)$, and so on, are the first, second, etc. derivatives of $f(x)$ evaluated at $x = x_0$. Thus they are constants, independent of x, once x_0 is selected.

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x-x_0)^2}{2!} f''(x_0) + \ldots + \frac{(x-x_0)^n}{n!} f^{(n)}(x_0) + R_n(x, x_0). \tag{21}$$

The Weierstrass Theorem states that every function which is continuous on a closed interval can be approximated on that interval as closely as we please by a polynomial of sufficiently high degree.

The practical hope derived from these theorems is that even low degree polynomials, say quadratics and cubics, may give good approximations if the interval involved is not too large and the function is fairly smooth.

Two basic practical facts are: first, polynomial models have been used with much success by experimenters, with and without statistics; second, the estimation of unknown parameters in polynomial regression models by least squares leads to equations which are linear in the unknowns, and hence can be solved by more-or-less routine arithmetical operations.

An additional robust feature of regression models is that if inadequate they are to some extent self-revealing. Thus, it is well known from least squares theory that, with moderately reasonable behavior of "errors", the residual sum of squares after fitting a given regression model will, when divided by a suitable factor, often called the residual "degrees of freedom", be an estimate of the residual variation -- if the model fitted was appropriate. Thus if through replication or other information we have independent knowledge of the magnitude of the error variance, then a check can be made on the model used. This procedure is, in fact, properly regarded as an analysis of variance technique and is an important part of the use of regression models. Thus in the absence of knowledge of functional relations among quantitative variables, polynomial regression models constitute moderately robust vehicles for organizing, analyzing, and summarizing experimental data. There are, however, a number of possible snags which must be kept in mind.

Item 1: However good the fit of a regression model over the range of variables for which data are available, extrapolation beyond the observed range is fraught with hazard, unless theory or other experiments give clear indication of the functional form in the region of extrapolation.

Item 2: Despite the self-checking of the regression model, even interpolation must be done carefully in that representation of the model may be systematically bad over some regions. This aspect can be studied and guarded against to some extent by the

computation and plotting of residuals. Happily, this practice is being recommended increasingly these days.

Item 3: The statistical methods for fitting regression models have good properties when the independent variables are free of important random errors. In many practical cases the independent variables are not free of errors. Just how misleading this can be is a topic which still needs much investigation.

Item 4: An open question always exists as to the degree of the polynomial which should be fitted. This problem becomes especially important when no reliable independent estimate of errors exists. There are real dangers in overfitting or underfitting and thereby assessing the importance of various factors improperly.

Item 5: When two or more variables are involved, it will often be sensible, in principle, to examine several regression models simultaneously, as for example those given in (22).

$$
\begin{aligned}
&y = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{20}x_1^2 + a_{02}x_2 , \\
&y = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{11}x_1x_2 + a_{02}x_2^2 , \\
&y = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{20}x_1^2 + a_{11}x_1x_2 , \\
&y = a_{00} + a_{01}x_2 + a_{11}x_1x_2 + a_{11}x_1^2 , \\
&\text{etc.}
\end{aligned}
\tag{22}
$$

The computing labor involved will usually present a formidable barrier, though automatic high-speed machines should eventually overcome this.

Item 6: The use of a standard shotgun technique such as fitting polynomial models can discourage careful thinking about specific situations by providing an easy but mediocre substitute. There is a long run danger of replacing insight by formalized numerical computations.

Item 7: The use of regression models is usually predicated on the assumptions that the factor levels involved are completely identified by the numbers associated with them. This may not be valid. For example, if the deformation behavior of a substance is being studied at say 5 levels of pressure the relevant features of the levels may be not only the final pressure but also the rate of pressure increase, the mechanism of pressure application, temperature increases due to the pressure, and so on. The factor levels are then quite definitely distinguishable but not so precisely identifiable by a single number. In such cases the results of analysis by a classification model could differ importantly from

those from a regression model. The analysis based on the classification model would be less specific, but usually more robust, than the regression model analysis.

Item 8: Regression models are usually frankly empirical. They are not, in general, based on broad theories which may be useful in wider circumstances. Conversely, the unthinking use of regression models does little to encourage the construction of broad scientific theories.

The listing of these items is not intended to disparage regression models nor to discourage their use; rather, it is hoped that, as in the case of classification models, the tool may be employed more efficiently if its weaknesses are recognized.

Relations Between Classification and Regression Models. In a side by side discussion of both classification and regression models there are implicit two challenges. One is the question of the relationships, if any, between these two types of models. We have, after all, claimed considerable generality for both types. Thus, despite their different justifications and interpretations they must relate in some systematic way.

The second challenge is, of course, what to do about combined qualitative and quantitative factors. Suppose, for example, we have an experiment with one qualitative and one quantitative factor. One simple answer is use a distinct regression model for the quantitative factor for every level of the qualitative factor. This may not be a bad procedure and sometimes will have much to recommend it, but in general seems an inadequate substitute. The remainder of the paper is devoted to a brief and rather superficial consideration of these two related questions.

Let us fix our attention on two factors $\mathcal{A}$ and $\mathcal{B}$ having A and B levels respectively. We know that we can, under quite general conditions, write a population classification model as in expression (9) and develop it into a statistical model for the observations as sketchily indicated in expressions (15) and (16). If the levels of factors $\mathcal{A}$ and $\mathcal{B}$ are quantitatively identified by the variables u and v, then usually we can also write a polynomial regression model such as (23).

$$
\begin{aligned}
y_{ij} = {} & \alpha_{oo} + \alpha_1 u_i + \alpha_2 u_i^2 + \alpha_3 u_i^3 + \cdots \\
& \beta_1 v_j + \beta_2 v_j^2 + \beta_3 v_j^3 + \cdots \\
& \gamma_{11} u_i v_j + \gamma_{12} u_i v_j^2 + \gamma_{21} u_i^2 v_j + \cdots
\end{aligned}
\tag{23}
$$

For a given range of levels in the populations of levels of factors $\mathcal{A}$ and $\mathcal{B}$, we can now inquire what are the relations between the components of the population classification and regression models? Straightforward algebra leads us to the results given in expression (24).

$$\mu = \alpha_{oo} + \alpha_1\bar{u} + \alpha_2\overline{u^2} + \alpha_3\overline{u^3} + \ldots$$

$$+\beta_1\bar{v} + \beta_2\overline{v^2} + \beta_3\overline{v^3} + \ldots$$

$$+\gamma_{11}\overline{uv} + \gamma_{12}\overline{u\bar{v}^2} + \gamma_{21}\overline{u^2}\bar{v} + \ldots.$$

$$a_i = (\alpha_1 + \gamma_{11}\bar{v} + \gamma_{12}\overline{v^2} + \gamma_{13}\overline{v^3} + \ldots)(u_i - \bar{u})$$

$$+(\alpha_2 + \gamma_{21}\bar{v} + \gamma_{22}\overline{v^2} + \gamma_{23}\overline{v^3} + \ldots)(u_i^2 - \overline{u^2})$$

$$+(\alpha_3 + \gamma_{31}\bar{v} + \gamma_{32}\overline{v^2} + \ldots)(u_i^3 - \overline{u^3})$$

$$+ \ldots,$$

(24)

$$b_j = (\beta_1 + \gamma_{11}\bar{u} + \gamma_{21}\overline{u^2} + \ldots)(v_j - \bar{v})$$

$$+ (\beta_2 + \gamma_{12}\bar{u} + \gamma_{22}\overline{u^2} + \ldots)(v_j^2 - \overline{v^2})$$

$$+ (\beta_3 + \gamma_{13}\bar{u} + \gamma_{23}\overline{u^2} + \ldots)(v_j^3 - \overline{v^3})$$

$$+ \ldots,$$

$$(ab)_{ij} = \gamma_{11}(u_i - \bar{u})(v_j - \bar{v})$$

$$+\gamma_{12}(u_i - \bar{u})(v_j^2 - \overline{v^2}) + \gamma_{21}(u_i^2 - \overline{u^2})(v_j - \bar{v})$$

$$+\gamma_{22}(u_i^2 - \overline{u^2})(v_j^2 - \overline{v^2}) + \ldots.$$

Definitions: $\bar{u} = \frac{1}{A}\sum_1 u_i$; $\overline{u^2} = \frac{1}{A}\sum_1 u_i^2$; etc.

It can be seen from (24) how the definition of the main effects of factor $\mathcal{A}$ depends on the "interaction coefficients" – the γ's – and on the levels of factor $\mathcal{B}$ involved through the means of v, v^2, v^3, etc. The difference between two main effects of levels of factor $\mathcal{A}$ is given in (25).

$$a_i - a_{i'} = (\alpha_1 + \gamma_{11}\bar{v} + \gamma_{12}\overline{v^2} + \ldots)(u_i - u_{i'})$$

$$+ (\alpha_2 + \gamma_{21}\bar{v} + \gamma_{22}\overline{v^2} + \ldots)(u_i^2 - u_{i'}^2)$$

$$+ \ldots$$

(25)

We see that this difference still depends importantly, in general, on the values of $\bar{v}$, $\overline{v^2}$, etc., and hence on just what levels of factor $\mathcal{B}$ are

involved. The difference between two A main effects does not, however, depend on what other levels of factor A are involved in the experimental situation.

The interactions $(ab)_{ij}$ will evidently be negligible if and only if the coefficients $\gamma_{11}, \gamma_{12}, \gamma_{21}$, etc. are all negligible. If this is so, then we see that the definitions of the a_i becomes independent of $\bar{v}$, $\overline{v^2}$, $\overline{v^3}$, etc. In other words, if the interaction coefficients, the γ's, are negligible, the meaning of the main effects of factor A become independent of which levels of factor B are involved in the experimental situation.

It is tempting to think that if the main effects of factor A are small, then the regression coefficients α_1, α_2, etc., will be small. The relations of expression (24) show that this is not at all necessarily so.

There is some suggestion on how to handle the combined qualitative-quantitative case in expression (24). Suppose, for simplicity, that a reasonable polynomial model would be as given in (26).

$$(26) \qquad Y_{ij} = \alpha_0 + \alpha_1 u_i + \alpha_2 u_i^2 + \beta_1 v_j + \beta_2 v_j^2 + \gamma_{11} u_i v_j .$$

If the levels of factor A are not quantitatively identified, then the u_i values are unknown. If we superimpose on (26) the appropriate classification population model, we obtain (27).

$$(27) \qquad Y_{ij} = \mu + a_i + \lambda_i (v_j - \bar{v}) + \beta_2 (v_j^2 - \overline{v^2}) ,$$

$$\text{Definition:} \quad \lambda_i = (\beta_1 + \gamma_{11} u_i) .$$

In this model the unknown parameters are as listed in (28).

$$(28) \qquad \mu, \{a_i\}, \{\lambda_i\}, \beta_2 ; \quad \sum a_i = 0 .$$

This crossed population model can be carried forward into a statistical model for the observations. The structure of the least squares estimates of the parameters is given in (29).

The usefulness of such models will have to be learned by field trial, as well as from further theoretical study.

$$(29) \qquad \begin{aligned} \hat{\mu} &= \bar{x}.. , \\ \hat{a}_i &= \bar{x}_{i.} - \bar{x}.. , \\ \hat{\beta}_2 &= \frac{S_{vv} S_{xv}^2 - S_{xv} S_{vv}^2}{S_{v^2v^2} S_{vv} - S_{vv}^2} , \\ \hat{\lambda}_i &= \frac{S_{xv}^i - \hat{\beta}_2 S_{vv}^2}{S_{vv}} = \frac{S_{xv^2v^2v} - S_{xv^2}}{S_{vv} S_{v^2v^2} - S_{vv^2}^2} . \end{aligned}$$

(29 cont.)

Definitions: $S_{vv} = \sum_j (v_j - \bar{v})^2$,

$S_{xv} = \sum_j (x_{.j} - x_{..})(v_j - \bar{v})$,

$S_{vv}{}^2 = \sum_j (v_j - \bar{v})v_j^2$,

$S_{v^2v^2} = \sum_j (v_j^2 - \overline{v^2})v_j^2$,

$S_{xv}{}^2 = \sum_j (x_{.j} - x_{..})v_j^2$,

$S_{xv}^i = \sum_j x_{ij}(v_j - \bar{v})$.

References

1. G. E. P. Box and S. L. Anderson. "Permutation theory in the derivation of robust criteria and the study of departures from assumption." Jour. Roy. Stat. Soc., 17, 1-34, 1955.

2. J. Cornfield and J. W. Tukey. "Average values of mean squares in factorials." Annals Math. Stat., 27, 907-949, 1956.

3. R. A. Fisher, "The arrangements of field experiments." J. Min. Agr. Eng., 33, 503-513, 1926.

4. ——— The Design of Experiments. Oliver and Boyd, London, 1935, 1953.

5. H. Scheffe. "Alternative models for the analysis of variance." Annals Math. Stat., 27, 251-271, 1956.

6. M. B. Wilk and O. Kempthorne. "Fixed, mixed and random models." Jour. Amer. Stat. Assoc., 50, 1144, 1167, 1955.

7. ——— "Some aspects of the analysis of factorial experiments in a completely randomized design." Annals Math. Stat., 27, 950-985, 1956.