REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE October 1960	3. REPORT TYPE AND DATES COVERED

4. TITLE AND SUBTITLE
Proceedings of the Fifth Conference on the Design of Experiments in Army Research Developments and Testing

5. FUNDING NUMBERS

6. AUTHOR(S)
Not Available

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Army Mathematics Advisory Panel

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

ARO-OORR 60-2

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This is a Technical report resulting from the Proceedings of the Fifth Conference on the Design of Experiments in Army Research Developments and Testing.

14. SUBJECT TERMS

15. NUMBER OF PAGES
429

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	**UNCLASSIFIED**	**UNCLASSIFIED**	**UL**

OORR 60-2

Office of Ordnance Research

PROCEEDINGS OF THE FIFTH CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH DEVELOPMENTS AND TESTING

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

This document contains
blank pages that were
not filmed.

OFFICE OF ORDNANCE RESEARCH
Report No. 60-2
October 1960

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

Sponsored by the Army Mathematics Steering Committee
conducted at
The U. S. Army Biological Warfare Laboratories
Fort Detrick, Frederick, Maryland
4-6 November 1959

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

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont'd)

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

TABLE OF CONTENTS (Cont'd)

---

* This paper was presented by title. It does not appear in this technical manual.

** This paper was presented by title.

FOREWORD

The present series of Conferences on the Design of Experiments are sponsored by the Army Mathematics Steering Committee (AMSC). The first three annual meetings were held at the Diamond Ordnance Fuze Laboratories and the National Bureau of Standards in Washington, D. C., and the fourth meeting was conducted at the Quartermaster Research and Engineering Center at Natick, Massachusetts. At its April 1959 meeting the AMSC accepted the invitation, issued by Dr. Clifford J. Maloney on behalf of the U. S. Army Biological Warfare Laboratories, to hold the Fifth Conference on the Design of Experiments at Fort Detrick, Maryland.

The purpose of these Conferences is to afford Army scientific and technological experts an opportunity to exchange views and experiences on problems of designing experiments in research, development and testing, and to learn about new developments in the field from experts in the design of experiments. The success of these Conferences has been due, in large measure, to the interaction and cooperation of these two groups of experts.

The Fifth Conference was attended by 169 registrants and participants from 60 organizations outside of the Biological Laboratories. In addition the host had 71 of its personnel present. Speakers and panelists came from Advanced Research Projects Agency, Bureau of Ships of the Department of the Navy, Mayo Clinic, National Bureau of Standards, Princeton University, RCA Missile Test Project, University of California, University of Georgia, University of Michigan, University of Toronto, Virginia Polytechnic Institute and 15 Army facilities.

This volume of the Proceedings contains 27 of the papers which were presented at the conference. In addition, it contains one of the two articles that were presented by title. 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 the many speakers and other research workers who participated in the meeting; to Colonel Clyde Westbrook, Commanding Officer of the U. S. Army Biological Warfare Laboratories, for making available the excellent facilities of his organization for the Conference; and to Dr. Clifford J. Maloney who handled the details of the local arrangements for the meeting, which included interesting tours of the Laboratories and of nearby Civil War battlefields such as Gettysburg, Antietam and Harper's Ferry.

Finally, the Chairman wishes to express his appreciation to his Advisory Committee, F. G. Dressel (Secretary), Frank E. Grubbs, Boyd Harshbarger, Clifford J. Maloney, and W. J. Youden for their help in organizing the program of the Conference.

S. S. WILKS
Professor of Mathematics
Princeton University

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

Wednesday AM
4 November

0830 - 0900 REGISTRATION: Post Theater

0900 - 1145 GENERAL SESSION I: Post Theater

Chairman
Dr. I. R. Hershner, Jr., Army Research Office; Office, Chief of Research and Development.

0900 - 0910 Welcome
Col. Donald G. Grothaus, Commanding Officer, Fort Detrick.

0910 - 0925 Introductory Remarks
Dr. Leroy D. Fothergill, Scientific Advisor, Fort Detrick.

0925 - 0930 Announcements
Dr. Morton Reitman, Technical Information Div., Fort Detrick.

0930 - 1030 The Method of Paired Comparisons
Dr. H. A. David, Virginia Polytechnic Institute.

1030 - 1045 Break

1045 - 1145 The Measure of Death
Dr. Joseph Berkson, Mayo Clinic.

1200 - 1300 LUNCH: Officers' Club

Wednesday PM
4 November

1300 - 1700 TOUR: Battlefield tour of Gettysburg or Antietam and Harpers Ferry. (Buses will depart from the Officers' Club)
Harpers Ferry National Monument:
Superintendent, Mr. Frank H. Anderson
Historian, Mr. Charles Snell
Antietam National Battlefield Site:
Superintendent, Mr. H. W. Doust
Historian, Mr. R. L. Lagemann
Gettysburg National Military Park:
Superintendent, Mr. James Myers
Historian, Mr. Frederick Tilberg

Wednesday PM (Cont'd)

1800 - 1900 SOCIAL HOUR: Officers' Club

1900 - 2000 DINNER: Officers' Club

2000 - 2200 GENERAL SESSION II: Officers' Club

Chairman: Dr. Clifford J. Maloney, Chief, Mathematics Division, Fort Detrick.

2000 - 2100 The Army Research and Development Program as it Relates to the Civil Economy
Dr. Richard Weiss, Army Research Office, Arlington Hall Station, Va.

2100 - 2200 Prediction of the Reliability of Complex Systems
Dr. Nicholas E. Golovin, Director, Technical Operations Division, Advanced Research Projects Agency.

There will be one Clinical and three Technical Sessions conducted Thursday morning. Technical Session I and Clinical Session A will both be held from 0830 - 1040. From 1100 - 1230 Technical Sessions II and III will be running concurrently. The security classification of the first paper in Technical Session III is CONFIDENTIAL. No clearances are required for any of the other papers on this program.

Thursday AM
5 November

0830 - 1040 TECHNICAL SESSION I: Post Theater

Chairman: Mr. Elwood K. Wolfe, Technical Evaluation Division, Fort Detrick.

0830 - 0910 On the Repeated-Measurements Design in Biological Experiments
Ardie Lubin, Department of Clinical and Social Psychology, Division of Neuropsychiatry, Walter Reed Institute of Research, WRAMC.

0910 - 0950 Design of Experiments Using Germfree Animals
Stanley M. Levenson, Ole J. Malm, and Captain Richard E. Horowitz, Department of Surgical Metabolism and Physiology, and the Department of Germfree Research, Walter Reed Army Institute of Research, WRAMC.

0950 - 1005 Break

TECHNICAL SESSION I: (Cont'd)

1005 - 1040 The Development of Parameters for Determining the Resistance of Selected Missile Components to Microbiological Deterioration
C. Bruce Lee, Physical Sciences Laboratory, Research and Engineering Directorate, Ordnance Tank-Automotive Command.

1040 - 1100 Break

0830 - 1040 CLINICAL SESSION A: Class Room, Bldg. T-833

Chairman: Mr. O. P. Bruno, Surveillance Branch, Weapon Systems Laboratory, Ballistic Research Laboratories.

Panel Members:
Besse Day, Bureau of Ships, Dept. of the Navy
Frank Grubbs, Weapon Systems Laboratory, Ballistic Research Laboratories
Boyd Harshbarger, Virginia Polytechnic Institute
G. M. Jenkins, Princeton University
R. G. D. Steel, Mathematics Research Center
S. S. Wilks, Princeton University

0830 - 0905 Design of Environmental Experiments for Reliability Prediction
A. Bulfinch, Nuclear and Advanced Systems Laboratory, Feltman Research and Engineering Laboratory, Picatinny Arsenal.

0905 - 0940 Multidimensional Staircase Designs for Reliability Studies
David R. Howes, U. S. Army Chemical Corps Engineering Command

0940 - 0955 Break

0955 - 1040 Approach to Development Policies Concerning Scheduled and Unscheduled Maintenance:
Walton M. Hancock and Randall Cline, The University of Michigan, Willow Run Laboratories, Operations Research Department.

1040 - 1100 Break

1100 - 1230 TECHNICAL SESSION II: Post Theater

Chairman: Dr. Robert M. Thrall, The University of Michigan

TECHNICAL SESSION II: (Cont'd)

1100 - 1140 Statistical Analysis of Various Parameters of Burning Characteristics of Flare Systems
Bossie Jackson, Pyrotechnics Laboratory, Feltman Research and Engineering Laboratory, Picatinny Arsenal.

1140 - 1150 Break

1150 - 1230 A Statistical Evaluation of the Pyrotechnic Electrostatic Sensitivity Tester
Everett D. Crane, Pyrotechnic Laboratory, Feltman Research and Engineering Laboratory, Picatinny Arsenal.

1100 - 1230 TECHNICAL SESSION III: Conference Room, Bldg. P-560

Chairman: Mr. B. A. Howard, Jr., Headquarters Ordnance Weapons Command

1100 - 1145 Dispersion Strengthening Analysis of Cermets
John M. Woulbroun, Sintered Metals and Ceramics Branch, Rodman Laboratory, Watertown Arsenal.

1145 - 1155 Break

1155 - 1230 Experimental Determination of "Best" Component Levels in Thermal Power Supplies (U). (Contents of talk CONFIDENTIAL)
Sheldon G. Levin, Diamond Ordnance Fuze Laboratories.

1230 - 1330 LUNCH: Picnic Lunch, Flair Armory. Buses to the armory will leave from the Officers' Club immediately following Technical Session III. There will be movies following lunch. Afterwards, buses will take you to the departure point for the walking tour.

Thursday PM
5 November

1330 - 1700 TOUR: Walking tour of Frederick, Maryland

1800 - 1900 DINNER: Peter Pan Restaurant. Buses to the restaurant will leave from the Francis Scott Key Hotel at 1730.

1900 - 2115 GENERAL SESSION III: Peter Pan

Chairman: Dr. S. S. Wilks, Princeton University

1900 - 2000 Medical Health Statistics
Dr. Wilford J. Dixon, University of California Medical Center.

GENERAL SESSION III: (Cont'd)

2000 - 2015 Break

2015 - 2115 Sampling in Biological Populations
Dr. D. B. DeLury, University of Toronto.

Friday AM
6 November

0830 - 1040 TECHNICAL SESSION IV: Post Theater

Chairman: Mr. John P. Purtell, Research Branch, Watervliet Arsenal.

0830 - 0910 The Application of Fractional Factorials in Missile Test Programs
Paul C. Cox, Reliability and Statistics Office, Ordnance Mission, White Sands Missile Range.

0910 - 0940 The Design and Re-design of an Experiment
C. W. Mullis, Plans Branch, Integrated Range Mission, White Sands Missile Range.

0940 - 0950 Break

0950 - 1015 On a Problem of Misclassification:
A. C. Cohen, Jr., The University of Georgia

1015 - 1040 Detecting and Quantifying Guess Responses in the Rating of Statements by a Method of Successive Intervals
Lee E. Paul, Methods and Systems Engineering Branch, Quartermaster R and E Field Evaluation Agency.

0830 - 1040 CLINICAL SESSION B: Class Room, Bldg. T-833.

Chairman: Mr. John Kosar, Missile Warheads and Special Projects Laboratory, FREL, Picatinny Arsenal.

Panel Members:
H. A. Daivd, Virginia Polytechnic Institute
D. B. DeLury, University of Toronto
W. J. Dixon, University of Cal. Medical Center
W. D. Foster, Fort Detrick
J. S. Hunter, Mathematics Research Center, U. S. Army
W. J. Youden, National Bureau of Standards

CLINICAL SESSION B: (Cont'd)

0830 - 0900 Design for Estimation by Covariance Techniques:
Morris Rhian, Aerobiology Division, U. S. Army Biological Warfare Laboratories.

0900 - 0920 Design of an Experiment to Evaluate a Bio-assay with Non-parallel Slopes
Albert L. Fernelius, Process Research Division, U. S. Army Biological Warfare Laboratories.

0920 - 0930 Break

0930 - 1010 The ORO Aircraft Vulnerability Experiment
Bruce Taylor, Operations Research Office, The Johns Hopkins University

1010 - 1040 Operational Hit Probabilities of Experimental Anti-tank Weapons
J. D. Reed, R. E. Tiller, and J. P. Young, Operations Research Office, The Johns Hopkins University.

1055 - 1230 TECHNICAL SESSION V: Post Theater

Chairman: Dr. H. Leon Harter, Wright Air Development Center, Wright Patterson Air Force Base.

1055 - 1135 Elimination of Bias Introduced by Transformation of Variables
Jerzy Neyman and Elizabeth L. Scott, Statistical Laboratory, University of California, Berkeley.

1135 - 1145 Break

1145 - 1230 Mathematical and Statistical Principles Underlying Chemical Corps Inspection Procedures for Product Verification
Henry Ellner and Joseph Mandelson, Materiel Command at the Army Chemical Center.

1055 - 1230 TECHNICAL SESSION VI: Class Room, Bldg. T-833.

Chairman: Mr. Abraham Golub, Support Weapons Evaluation Branch, Weapon Systems Laboratory, Ballistic Research Laboratories.

1055 - 1140 Measuring a Complex Field Operation
K. L. Yudowitch, Operations Research Office, The Johns Hopkins University

1140 - 1150 Break

TECHNICAL SESSION VI: (Cont'd)

1150 - 1230 The Conduct of Military Field Research on a Shoe-String
A. J. Eckles, III, and R. E. Zimmerman, Operations Research Office, The Johns Hopkins University.

1230 - 1330 LUNCH: Optional

Friday PM
6 November

1330 - 1500 TOUR: A conducted tour of the Fort Detrick Laboratories to start from the Officers' Club.

SUPPLEMENTARY PROGRAM

We are sorry that time did not permit the scheduling of the following two papers. These authors, as well as all speakers on this program, are urged to submit manuscripts of their papers so that a complete and interesting technical manual can be published. A copy of these Proceedings will be sent to each attendee of this conference.

Sample Order Statistics of the Circular Normal Distribution
Helen J. Coon, Weapon Systems Laboratory, Ballistic Research Laboratories.

Determination of Systematic Errors in Tracking Radar
Victor B. Kovac, RCA Missile Test Project, Patrick Air Force Base.

THE METHOD OF PAIRED COMPARISONS

H. A. David
Virginia Polytechnic Institute

INTRODUCTION. In a paired-comparison experiment objects or "stimuli" are presented in pairs to a panel of judges who act independently. The basic experimental unit is the comparison of two objects, A and B, by a single judge who, in the simplest situation, must state which one he prefers. One may also allow the judge the third alternative of declaring a tie. A further generalization would be to give the judge a scale of preferences; for example, a seven-point scale reading "strong preference for item A," "preference for A," "slight preference for A," "no preference," "slight preference for B," "preference for B," "strong preference for B." These preferences may be scored by assigning object A the score i (i = 3, 2, 1, 0, -1, -2, -3) and B the score -i. A slightly different scoring system prevails in a widely publicized form of paired comparison such as we have recently been witnessing in the series between the Dodgers and the White Sox where each game corresponds to one paired comparison and the series to several repetitions.

The comparison of A and B may be made by all the judges. If more than 2 objects are to be compared it is still possible to arrange that every judge makes every possible paired comparison either once or several times. This situation may be called a balanced paired-comparison experiment and corresponds in the language of sport to a Round Robin tournament; the roles of the players in the tournament being analogous to those of the objects in the paired-comparison experiment. If we have t objects and n judges the number of paired comparisons will be $\frac{1}{2}rn\ t\ (t-1)$, where r is the number of times a particular judge makes a particular paired comparison or in other words, the number of replications of a simple Round Robin tournament.

The method of paired comparisons is used primarily in cases when the objects to be compared can be judged only subjectively; that is to say, when it is impossible or impracticable to make relevant measurements in order to decide which of two objects is preferable. As may be inferred, paired comparisons are widely employed by psychometricians, and the method was indeed first introduced by Thurstone (1927). Most frequent applications have been to taste testing, color comparisons, personnel rating, and generally to all forms of preference testing. Of course, there are other methods of sensory discrimination and it is not proposed to enter into a detailed discussion of the individual merits of these methods, particularly as a number of summary accounts have recently been given [Jones and Bock (1957), Torgerson (1958) and Bliss (1959)]. The method of paired comparisons is sometimes the only practicable experimental procedure as in testing various brands of razors where two razors can be compared on a man's two cheeks. Sometimes it may be possible for a judge to compare several objects at the same time and if this can easily be done it would indeed be preferable for the judge to assign ranks to all these objects. However, when differences between objects are small it is advantageous to make the comparison between two of them as free as

possible from any extraneous influences such as may be provided by taking into consideration other objects at the same time. Thus the method of paired comparisons will be used in cases where a fine judgment is needed. Again, in taste testing it is often not possible for a judge to cope with more than two tastes, and the introduction of a third taste may be thoroughly confusing.

When both paired comparisons and ranking are possible procedures in arranging several objects in order of preference, ranking will certainly be the speedier. On the other hand, the method of paired comparisons makes it possible for the judge to contradict himself; for example, he may prefer A over B, B over C, and yet C over A. This situation is certainly not impossible and has been called a circular triad by Kendall. An extreme example is provided by the game of stone, scissors, and paper. It is clear that if one judge is guilty of considerably more circular triads than another, then he is a less consistent judge. We have, therefore, a basis for a method for selecting good judges. The explanation of a circular triad may be that the judge is essentially guessing or it may be that in making the three comparisons he changes the criterion on which he bases his judgment. Putting it in different words, the preference scale may well not be uni-dimensional. A preference may be based on a number of characteristics of the objects and presumably these characteristics are weighted in some way in the judge's mind before he comes to a decision. The weights assigned may well vary from comparison to comparison for an inexperienced judge.

In the remainder of this paper we shall consider a number of points arising in the design and analysis of paired-comparison experiments, with special emphasis on some work recently done at the Virginia Polytechnic Institute.

THE DESIGN OF PAIRED-COMPARISON EXPERIMENTS. In the language of the design of experiments a Round Robin tournament is simply a balanced incomplete block design with judges corresponding to replications and with block size 2. Questions of design become more difficult when it is not feasible for every judge to make all possible comparisons. A very considerable degree of balance can sometimes be retained by what Bose (1956) has termed "linked paired comparison designs." An example of such a design is given in Table 1. Even more balance could be obtained if it is important to eliminate effects due to order of presentation within a pair. Related problems are discussed by Kendall (1955). Simpler but less well balanced methods of partial pairing had previously been developed by McCormick and Bachus (1952) in connection with the rating of a large number of employees.

It is a well-established dogma of experimental design that an experiment should contain a large degree of balance. There are, however, situations when balance is a doubtful asset. If we are interested in discovering the best of a number of treatments it is intuitively more reasonable to proceed sequentially - if this is practicable - in a fashion which will result in more intensive testing of those treatments most successful in the early stages of the experiment. Recalling that

a balanced paired-comparison experiment is equivalent to a Round Robin tournament we are led to consider other types of tournaments such as the Knock-out which have as their aim picking the strongest of a group of players.

Consider a tournament of 4 players. A simple (i.e. unreplicated) Round Robin tournament requires 6 games, as do two replications of a Knock-out tournament. As a first step toward a wider comparison one may therefore investigate the effectiveness of these two tournaments in determining the best player. This may be done by assigning values to each π_{ij}, the probability that player i defeats player j, and finding the probability that the strongest player (the player for whom $\pi_{i\cdot}$ is largest) will win the tournament. In calculating this probability we average over all possible draws. The situation is unfortunately complicated by the possible need for play-offs if two or three players end up in the lead. In addition to the probability that the best player will win it is therefore advisable to take into consideration the expected number of games required to determine the winner. Both criteria are evaluated in [3] by enumeration of all possible outcomes of the tournament and determination of their probabilities. In a series of examples studied the Knock-out tournament does in fact emerge as superior on both counts in nearly all cases. Another type of tournament employs double elimination; that is, first round losers are paired off and a player is eliminated only after losing to two opponents. This turns out to be the best of three types of tournament. A variation of the Knock-out tournament, which in any match between 2 players requires not one but the best out of 3 games to determine the winner, has been suggested by Maurice (1958). It is not easily compared with the other tournaments, except on the basis of a cost function, as it tends to require more games in return for a higher probability of determining the best players.

The following is typical of the results obtained. With parameter values

$$\pi_{12} = .70,\ \pi_{13} = .76,\ \pi_{14} = .86,\ \pi_{23} = .75,\ \pi_{24} = .82,\ \pi_{34} = .72$$

the probability that player 1 (the strongest) will win and the corresponding expected number of games is

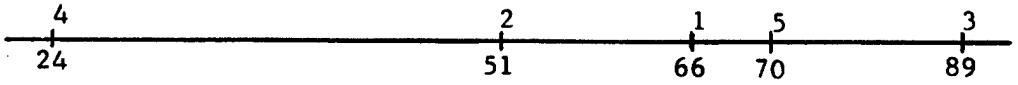
0.644, 6.62	for the Round Robin tournament,
0.656, 6.56	for the Knock-out tournament,
0.686, 6.43	for the Double Elimination tournament,
0.706, 7.08	for Maurice's tournament.

ESTIMATION PROCEDURES. We return now to a more detailed consideration of a balanced paired-comparison experiment in which each of n judges compares t objects r times. Further we suppose that each comparison results in a straight preference for one or the other object judged. The results for each judge can then be fully presented in the familiar

two-way table of 1's and 0's. In addition, the number of times each object is preferred to all others may be listed in a column of totals (the number of wins or score of each object, treatment, or player). If differences between judges can be assumed to be slight - and this can be tested - the n individual tables are conveniently amalgamated into a single summary table. For example, in the pairwise comparison of 5 brands of carbon paper by 30 secretaries (see Fleckenstein et al, 1958, and [2] for details) the following results, condensed from the original 7-point scale used, were obtained:

Brand	1	2	3	4	5	Total a_i
1	-	20	6	25	15	66
2	10	-	10	20	11	51
3	24	20	-	27	18	89
4	5	10	3	-	6	24
5	15	19	12	24	-	70
						300

Generally, the upshot of an experiment of this type has been the construction of a "response scale" in which the objects are appropriately spaced in increasing order of preference along a straight line. An obvious way of doing this is to use the total scores. Thus the results of the carbon paper experiment can represented as follows:



Here only the relative distances between scores are important.

This simple procedure may be regarded as a method of estimation. Let π_{ij} be the probabillty that in the comparison of objects i and j, i is preferred to j; and let

$$\pi_{i\cdot} = \sum_{\substack{j=1 \\ j \neq i}}^{t} \pi_{ij} \quad (= \sum_{j}{}' \pi_{ij}, \text{ say}).$$

Also let a_{ij} be the observed number of times that i is preferred to j, so that $a_i = \sum_j{}' a_{ij}$. Then clearly,

$p_{ij} = a_{ij}/n$ is an estimate of π_{ij}

$p_{i.} = a_i/[n(t-1)]$ is an estimate of $\pi_{i.}$.

It is surprising that this simple distribution-free method of estimation has not been more widely used. What has been usually done instead is to propose specific models giving the $\frac{1}{2}t(t-1)$ parameters π_{ij} in terms of t parameters (or t-1 if the origin of the scale is fixed).

Two cases have received special attention:

$$(1)\quad \pi_{ij} = \int_{-(S_i - S_j)}^{\infty} \frac{1}{\sqrt{(2\pi)}} e^{-\frac{1}{2}x^2} dx$$

(Thurstone, 1927; Mosteller, 1951), where the responses to the t objects are assumed to be equi-correlated normal variates with true means $S_i (i = 1,\ldots,t)$ and common variances.

$$(2)\quad \pi_{ij} = \pi_i/(\pi_i + \pi_j)$$

(Bradley and Terry, 1952), where the π_i are true "ratings" of the objects and satisfy $\pi_i \geq 0$, $\Sigma\pi_i = 1$.

If the models are appropraite they will generally lead to better scales than the simple scale above, which is however much more widely valid. (1) and (2) as well as two other scales have been compared by Jackson and Fleckenstein (1957) who found the four scales quite close in a color preference test.

SIGNIFICANCE TESTS. A question that arises naturally in the interpretation of a response scale, whatever its mode of derivation, is whether any differences between objects indicated by the scale are in fact statistically significant. Several methods of constructing over-all tests are available, that is tests of the null hypothesis H_o that all treatments are alike (in the responses they evoke). The simplest of these tests is to make use of the fact that

$$D = 4 \sum_{i=1}^{t} (a_i - \bar{a})^2 / (nt)$$

is, on H_o', distributed approximately as χ^2 with t-1 degrees of freedom. This is a special case of a more general test given by

Durbin (1951) and is equivalent to an older method based on counting the number of circular triads (Kendall and Smith, 1940). The goodness of the χ^2 approximation is examined in [2]. For the carbon paper experiment

$$D = 4 \text{ X } 2,\ 354/30 \text{ X } 5 = 62.77,$$

which is a highly significant value of χ^2 with 4 D.F.

This overall test leaves many questions unanswered, for example:

(1) If, prior to the experiment, one of the t objects in the paired-comparison experiment is of particular interest to the experimenter, how can he use the results to test whether this object is better (or worse) than, or different from, the average of all objects?

(2) If, before the experiment, there is a special interest in whether two specified objects produce different responses, how does one use the results of the full paired-comparison experiment to test for a difference?

(3) How does one test whether the object with the highest (lowest) score in the experiment is significantly better (worse) than the average?

(4) How does one order the t objects in a paired-comparison experiment into significantly different groups?

(5) How does one test whether the difference of two treatment scores which are chosen after the completion of the experiment is significant?

To answer questions (4) and (5) it is possible to adapt the well-known multiple comparison procedures due to Tukey and to Scheffé. This approach will not be treated here but is described in [2]. We now consider questions (1) - (3) in turn.

(1) Test of a pre-assigned object

Because of cost of some other characteristic of object r (O_r), $1 \leq r \leq t$, the experimenter may be particularly interested in knowing whether this object is better than average, that is, if

$$\pi_{r\cdot} = \sum_{j}{}' \pi_{rj}/(t-1) > \tfrac{1}{2}.$$

On H_o' the score a_r of O_r is a binomial variate with parameters $n(t-1)$, $\frac{1}{2}$. If a_r^o is the observed score of O_r the corresponding significance level is

$$\Pr(a_r \geq a_r^o \mid H_o') = 2^{-n(t-1)} \sum_{k=a_r^o}^{n(t-1)} \binom{n(t-1)}{k}.$$

Except in small experiments a normal approximation can be used to evaluate this probability.

In view of the generality of our model the point arises here and elsewhere that one may in fact be interested in testing not H_o' (the hypothesis that all objects are alike) but the more general null hypothesis

$$H_o: \ \pi_{i.} = \tfrac{1}{2} \qquad \text{all } i.$$

The two hypotheses are the same for the models of Thurstone and Bradley-Terry, and indeed for any linear model. It can be shown [5] that the above procedure is conservative under H_o; that is, the level of significance under H_o' is greater than under H_o.

(2) Tests of equality of two pre-assigned objects

Consider the case in which interest is expressed before the experiment in testing the difference between O_r and O_s. One therefore wishes to test H_o against one-sided or two-sided alternatives $\pi_{r.} > \pi_{s.}$ or $\pi_{r.} = \pi_{s.}$, respectively. This can be done by finding the distribution of $d = a_r - a_s$ under H_o'. Table 2 giving upper 5 and 1% points of d has been constructed from the exact distribution of d for small experiments and a normal approximation (with continuity correction) otherwise.

Illustration. In the carbon paper experiment brand 2 is more expensive than brand 4. Is it significantly better?

A one-sided test is required, say at the 5% level. We have $d = a_2 - a_4 = 51 - 24 = 27$. Also

$$1.64\sqrt{nt/2} + 0.5 = 1.64\sqrt{75} + 0.5 = 14.7$$

giving $d_c = 15$. Since $d \geq d_c$ we may declare brand 2 superior to brand 4.

(3) Test of the highest score

After running a paired-comparison experiment, the experimenter may wish to know whether the object with the highest score (a_{max}, say) is significantly better than average.

Let A_i be the event $a_i \geq m \left[0 \leq m \leq n(t-1)\right]$. Then by the principle of inclusion and exclusion

$$\Pr(a_{max} \geq m) = \Pr\left(\sum_{i=1}^{t} A_i\right)$$

$$= \sum_{j=1}^{t} (-1)^{j-1} \binom{t}{j} \Pr(A_1 A_2 \ldots A_j).$$

For small experiments it is possible to evaluate this probability exactly and tables are given in [1] for n = 1. In other cases it is often adequate to use the first term in the sum, viz.,

$$t \Pr(a_i \geq m) = t\, 2^{-n(t-1)} \sum_{k=m}^{n(t-1)} \binom{n(t-1)}{k}$$

as an approximation to $\Pr(a_{max} \geq m)$; it is, of course, also an upper bound. To test the significance of a_{max} approximately at level α one chooses as the critical value that positive integer m, say m_β, for which

$$t \Pr(a_i \geq m_\beta \mid H_o') = \beta \leq \alpha \leq t \Pr(a_i \geq m_\beta - 1 \mid H_o').$$

If $a_{max} \geq m_\beta$ one concludes that the object with score a_{max} is better than average at the 5% level of significance.

Illustration. In the carbon paper experiment brand 3 obtained the highest score: $a_{max} = 89$. To test whether this is significant at the 5% level we note from tables (e.g. Harvard Univ., 1955) that for sample size $n(t - 1) = 120$ and $p = \frac{1}{2}$

$$5 \Pr(a_i \geq 74) = 0.033$$

and $$5 \Pr(a_i \geq 73) > 0.05.$$

Thus $\beta = 0.033$ and $m_\beta = 74$.

Since $a_{max} > 74$, we conclude that brand 3 is significantly better than average.

THE TREATMENT OF TIES. In our discussion of estimation procedures and significance tests we have assumed that judges are not allowed to declare ties. This certainly simplifies the analysis but is frequently not desirable. Various methods for treating ties are in use: equal division among the tied objects, decision by the toss of a coin, and ignoring ties altogether. The last method has advantages in significance testing but is clearly unsuitable for the estimation of a response scale since it does not distinguish between results such as the following: A preferred 4 times, B once, no ties and A preferred 4 times, B once, 20 ties. The other two approaches may seem very plausible but if A is generally preferred to B it is likely, on the whole, to have had a slight edge on B even in those cases where the judge could reach no decision. The following model is proposed in [4]. Suppose that in the comparison of two objects O_i and O_j by a particular judge a response x_i is evoked by O_i and a response x_j by O_j. If $|x_i - x_j| \leq \tau$ the judge declares a

tie, if $x_i - x_j > \tau$ he prefers 0_i, if $x_j - x_i > \tau$ he prefers 0_j. Here the symbol τ denotes a sensory threshold. If $\tau = 0$ we are back in the situation where the probability of a tie is zero.

The model can be superposed on that of Thurstone and Mosteller. Least squares methods can then be used to estimate not only the mean responses S_i ($i = 1, \ldots, t$) but also the parameter τ (and possibly different τ's for different judges, a point which can be tested). Actually in [4] the differences $x_i - x_j$ were taken to follow a cosine law rather than the normal law of Thurstone. It should be noted that no splitting of the ties is actually made, the original observations being used in the analysis. The model has been found to give a satisfactory fit in the carbon paper experiment when the original 7-point scale is condensed into a 3-point scale.

Table 1

A linked paired comparison design for 5 treatments and 6 judges

Judge	Pairs assigned to a judge
a	(3, 5), (2, 4), (1, 3), (1, 4), (2, 5)
b	(2, 3), (3, 4), (1, 4), (1, 5), (2, 5)
c	(2, 3), (3, 5), (1, 2), (4, 5), (1, 4)
d	(3, 5), (1, 2), (3, 4), (2, 4), (1, 5)
e	(1, 2), (3, 4), (4, 5), (1, 3), (2, 5)
f	(2, 3), (4, 5), (2, 4), (1, 3), (1, 5)

$t = 5$ (no. of treatments or objects to be compared)

$n = 6$ (no. of judges)

$b = 10$ (no. of different pairs)

$r = 5$ (no. of pairs compared by each judge)

$k = 3$ (no. of times each pair is judged)

$\lambda = 2$ (no. of pairs compared in common by any two judges)

$\alpha = 2$ (no. of times each object is compared by each judge)

(From R. C. Bose (1956) with a slight change in notation)

Table 2

Critical values of d, the difference in scores of two pre-assigned objects (t = no. of objects, n = no. of replications)

Experiment Size		$\alpha = 0.01$		$\alpha = 0.05$	
		one-sided test	two-sided test	one-sided test	two-sided test
n	t	d'_c	d_c	d'_c	d_c
1	≤ 4	no significant values		no significant values	
1	5	4	none possible	4	4
1	6	5	5	4	4
1	7	5	5	4	5
1	8	5	6	4	5
1	9	6	6	4	5
1	10	6	7	5	5
1	11	6	7	5	6
1	12	7	7	5	6
1	13	7	7	5	6
1	14	7	8	5	6
1	15	7	8	5	6
1	16	7	8	6	6
2	3	no significant values		4	4
2	4	5	6	4	5
2	5	6	6	5	5
3	3	6	6	4	5
3	4	6	7	5	6
4	3	6	7	5	6
4	4	7	8	6	6
All larger values of n or t		d'_c = smallest integer $\geq 2.33\sqrt{\frac{1}{2}nt} + 0.5$	d_c = smallest integer $\geq 2.56\sqrt{\frac{1}{2}nt} + 0.5$	d'_c = smallest integer $\geq 1.64\sqrt{\frac{1}{2}nt} + 0.5$	d_c = smallest integer $\geq 1.96\sqrt{\frac{1}{2}nt} + 0.5$

REFERENCES

C. I. Bliss, "Some statistical aspects of preference and related tests," Proc. 4th Conference on Design of Expts. in Army Research Development and Testing (1958), pp. 249-271.

R. C. Bose, "Paired comparison designs for testing concordance between judges," Biometrika, Vol. 43 (1956), pp. 113-121.

R. A. Bradley and M. E. Terry, "The rank analysis of incomplete block designs. I. The method of paired comparisons," Biometrika, Vol. 39 (1952), pp. 324-345.

J. Durbin, "Incomplete blocks in ranking experiments," Brit. J. Psychol. (Statist. Sect.) Vol. 4 (1951), pp. 85-90.

Mary Fleckenstein, R. A. Freund and J. E. Jackson, "A paired comparison test of typewriter carbon papers," Tappi Vol. 41 (1958), pp. 128-130.

J. E. Jackson and Mary Fleckenstein, "An evaluation of some statistical techniques used in the analysis of paired comparison data," Biometrics, Vol. 13 (1957), pp. 51-64.

L. V. Jones and R. D. Bock, "Methodology of preference measurement," Final report, Quartermaster Food and Container Institute for the Armed Forces (1957), pp. 1-202.

M. G. Kendall, "Further contributions to the theory of paired comparisons," Biometrics, Vol. 11 (1955), pp. 43-62.

M. G. Kendall and B. Babington Smith, "On the method of paired comparisons," Biometrika, Vol. 31 (1940), pp. 324-345.

E. J. McCormick and J. A. Bachus, "Paired comparison ratings. I. the effect on ratings of reductions in the number of pairs," J. Appl. Psychol., Vol. 36 (1952), pp. 123-127.

Rita J. Maurice, "Selection of the population with the largest mean when comparisons can be made only in pairs," Biometrika, Vol. 45 (1958), pp. 581-586.

F. Mosteller, "Remarks on the method of paired comparisons: I. The least square solution assuming equal standard deviations and equal correlations," Psychometrika, Vol. 16 (1951), pp. 3-9.

L. L. Thurstone, "Psychophysical analysis," Amer. J. Psychol., Vol. 38 (1927), pp. 368-389.

W. S. Torgerson, Theory and methods of scaling, John Wiley and Sons (1958).

[1] H. A. David, "Tournaments and paired comparisons," Biometrika Vol. 46 (1959), pp. 139-149.

[2] T. H. Starks and H. A. David, "Significance tests in experiments involving paired comparisons," Tech. Rep. No. 41, Virginia Polytechnic Institute (1959).

[3] W. A. Glenn, "A comparison of the effectiveness of tournaments," Tech. Rep. No. 42, V.P.I. (1959).

[4] W. A. Glenn and H. A. David, "Ties in paired comparison experiments," Tech. Rep. No. 43, V.P.I. (1959).

[5] H. A. David, "A conservative property of binomial tests," Tech. Rep. No. 44, V.P.I. (1959).

MEASURE OF COMPETING EXPONENTIAL MORTALITY RISKS WITH ESPECIAL REFERENCE TO THE STUDY OF SMOKING AND LUNG CANCER

Joseph Berkson, M.D.
Mayo Clinic, Rochester, Minnesota

I shall consider the model of two competing risks in the sense of Neyman [10]; and to set out the problem, I take first a very simple example.

Two marksmen shoot at a range of targets, under conditions in which if a target is struck, it drops instantly from view so that it cannot be struck again. This provision is made because the striking of a target with a bullet is intended to represent the striking down of a man by death from disease. Let the striking rate of Marksman 1 (who may be taken to represent a specific disease), when he is firing alone, be q_1, and similarly let the rate when Marksman 2 is firing alone be q_2. The probability when one risk operates alone is called the "net" risk or rate, and is represented by lower case q; when it operates together with another risk, the resulting risk is called the "crude" risk or rate and is represented by capital Q.

Suppose N targets are exposed and Marksman 1 shoots first, followed by Marksman 2.

(1) Rate for 1 is $Q_1 = q_1$

(2) Rate for 2 is $Q_2 = (1 - q_1)\, q_2$

(3) Total rate is $Q_1 + Q_2 = q_1 + q_2 - q_1\, q_2$

Suppose, instead, Marksman 2 shoots first, followed by Marksman 1.

(4) Rate for 2 is $Q_2 = q_2$

(5) Rate for 1 is $Q_1 = (1 - q_2)\, q_1$

(6) Total rate is $Q = q_1 + q_2 - q_1\, q_2$

It is seen that the total crude rate, with both marksmen firing, is the same, whichever shoots first, and assuming independence of the net probabilities q_1 and q_2, this will be true in general. Regardless of the order of shooting, or whether the two marksmen shoot together, the total crude rate is given by (3) (6). This result would, of course, usually be derived as the complement of the product of the probabilities $p_1 = 1 - q_1$ and $p_2 = 1 - q_2$, of not being struck; that is as 1 minus the product of the survival rates.

If, from independent trial, we knew q_1, the net rate of Marksman 1, and had observed the result Q, the crude rate when both shot together, we could derive the net rate q_2 of Marksman 2 from (3):

$$q_2 = \frac{Q - q_1}{1 - q_1} \tag{7}$$

But suppose we did not know the net rate of either marksman, q_1 of q_2, but had observed the results of their shooting together, and could identify the number of targets struck by each, from the shape of the bullet hole or otherwise, so that we could determine the individual crude rates Q_1 and Q_2 -- still we could not determine the net rates q_1, q_2, from these data alone. We have seen that, with the same net rates q_1, q_2 operating, although the total crude rate Q is independent of the order of shooting, the individual crude rates Q_1, Q_2 depended on which marksman shot first. This problem of estimating a risk, from observations when another risk is operating with it, called "competing risks," by Neyman [10], arises in different contexts of many statistical problems.

In order to estimate the net q's from the observed crude Q's, something has to be known regarding the time relation of the risks. A simplifying assumption, which is frequently reasonable, is to suppose that each instantaneous risk, which is called the "force of mortality" in actuarial texts, is constant over the period of observation. If l_t is the number of survivors at time t, then

$$-\frac{dl_t}{l_t dt} = -\frac{d\ln l_t}{dt}$$

is the instantaneous risk. I will use β's to represent the instantaneous risks, and shift to the example of dealing with two mortality rates, q_1 the net mortality from some specified disease, and q_2 the net mortality rate from all other diseases than 1, taken together and considered as a single risk. Then the net probability of death from the respective causes at time $<t$ is given by

$$q_{1_t} = 1 - e^{-\beta_1 t} \tag{8}$$

$$q_{2_t} = 1 - e^{-\beta_2 t} \tag{9}$$

where β_1 is the instantaneous risk for net death risk 1, β_2 is the instantaneous risk for net risk 2, and t is the time measured from t = 0.

From (8), (9) we have the corresponding net probability of survival to time t

$$p_{1_t} = 1 - q_{1_t} = e^{-\beta_1 t} \tag{10}$$

$$p_{2_t} = 1 - q_{2_t} = e^{-\beta_2 t}. \tag{11}$$

The probability of survival to time t, with both risks operating together is the product of (10) (11)

$$P_t = e^{-(\beta_1 + \beta_2)t} = e^{-\beta t} \tag{12}$$

and the probability of dying at time $<t$ is

$$Q_t = 1 - P_t = 1 - e^{-\beta t} \tag{13}$$

where $\beta = \beta_1 + \beta_2$.

The formulas (10), (11), (12) represent "survival functions" in the context of actuarial discussions.

Without loss of generality, we can consider the period of observation as from t = 0 to t = 1.

The proportion of persons dying from cause 1 over the unit period, say a year, from t = 0 to t = 1 is the crude death rate from cause 1. It is

$$Q_1 = \int_0^1 e^{-\beta t} \beta_1 \, dt = \frac{\beta_1}{\beta} \left(1 - e^{-\beta}\right) = \frac{\beta_1}{\beta} \, Q \tag{14}$$

and similarly for cause 2

$$Q_2 = \frac{\beta_2}{\beta} \left(1 - e^{-\beta}\right) = \frac{\beta_2}{\beta} \, Q \tag{15}$$

and for total deaths from all causes

$$Q = Q_1 + Q_2 = 1 - e^{-\beta} \tag{16}$$

and the probability of survival to the end of the period is

$$P = 1 - Q = e^{-\beta}. \tag{17}$$

The net death rates over the unit period are

$$q_1 = 1 - p_1 = 1 - e^{-\beta_1} \tag{18}$$

$$q_2 = 1 - p_2 = 1 - e^{-\beta_2}. \tag{19}$$

Now, we observe the crude rates Q_1, Q_2 and $Q = Q_1 + Q_2$; we wish the net rates q_1, q_2. These can be derived directly from (14), (15), (18), (19), and are given by

$$\ln (1 - q_1) = -\beta_1 = \frac{Q_1}{Q} \ln (1 - Q) \tag{20}$$

$$\ln (1 - q_2) = -\beta_2 = \frac{Q_2}{Q} \ln (1 - Q) . \tag{21}$$

MAXIMUM LIKELIHOOD FREQUENCY ESTIMATE. The development of the formulas for obtaining the net rates q_1, q_2 just given in (20), (21) is what is sometimes called "deterministic." We simply solved algebraically for the q's, having written down the equations representing the assumptions. If we stop to think a moment, in making these solutions we said we knew the crude rates Q_1,Q_2. But how are we to know them? We assume that we have observed them -- the Q's represent the "observed" rates which are computed by dividing deaths by N. But from a statistical view, if the numbers N on which these observed rates are based are moderate or small, we do not "know" the Q's -- these are only estimates. I will now consider the problem from the stochastic view, and specifically will develop the maximum likelihood estimates and their variances.

N individuals are observed over the unit period from t = 0 to t = 1. We observed d deaths, d_1 from cause 1, d_2 from cause 2, and $s = N - d_1 - d_2$ survivors to the end of the period. First, it will be convenient to estimate $\beta = \beta_1 + \beta_2$. Since the crude probability of death is $(1 - e^{-\beta})$, and of survival it is $e^{-\beta}$, the probability of the sample is proportional to

$$\phi = (1 - e^{-\beta})^d \, e^{-\beta s} . \tag{22}$$

From (22) we derive the maximum likelihood estimate and its variance in the standard way.

$$\hat{\beta} = \ln (N/s) \tag{23}$$

$$\sigma^2_{\hat{\beta}} = \frac{1 - e^{-\beta}}{N \, e^{-\beta}} \tag{24}$$

To derive the estimates of β_1 and β_2 we write the probability of the sample in terms of d_1 and d_2. It will be remembered that the crude probability of death from cause 1 is $\frac{\beta_1}{\beta}(1 - e^{-\beta})$, and from cause 2 it is $\frac{\beta_2}{\beta}(1 - e^{-\beta})$, and the probability of survival to the end of the period is $e^{-\beta}$. The probability of the sample is then proportional to

$$(25) \qquad \mathfrak{p} = \left[\frac{\beta_1}{\beta}(1 - e^{-\beta})\right]^{d_1}\left[\frac{\beta_2}{\beta}(1 - e^{-\beta})\right]^{d_2} e^{-\beta s}$$

and from this we obtain

$$(26) \qquad \hat{\beta}_1 = \frac{d_1}{d} \ln (N/s) = \frac{d_1}{d}\hat{\beta}$$

$$(27) \qquad \hat{\beta}_2 = \frac{d_2}{d} \ln (N/s) = \frac{d_2}{d}\hat{\beta}$$

$$(28) \qquad \sigma^2_{\hat{\beta}_1} = 1/N\left[\frac{\beta_1 \beta_2}{(1 - e^{-\beta})} + \frac{\beta_1^2 (1 - e^{-\beta})}{\beta^2 e^{-\beta}}\right]$$

$$(29) \qquad \sigma^2_{\hat{\beta}_2} = 1/N\left[\frac{\beta_1 \beta_2}{(1 - e^{-\beta})} + \frac{\beta_2^2 (1 - e^{-\beta})}{\beta^2 e^{-\beta}}\right].$$

We obtain the estimate of the q's from the estimates of the β's by the corresponding relation to the parameters, for instance

$$q_1 = 1 - e^{-\beta_1}$$

$$\text{var. } q_1 = (1 - q_1)^2 \text{ var. } \beta_1 .$$

If these maximum likelihood estimates which I call the "frequency estimates" are examined, it will be found that, in effect, they are the same as the estimates derived on a deterministic basis, since in that case we take the crude probability Q as given by the corresponding observed relative frequency d/N. However, with the development of the maximum likelihood estimate, we have also the large sample variance.

MAXIMUM LIKELIHOOD TIME ESTIMATES. In developing the maximum likelihood frequency estimate as just completed, we took into account only the number of deaths from each cause in the unit period. We did not use any information on the times of the deaths. But if the survival

functions are of assumed form, these times should help us estimate the parameters β. I will now develop the maximum likelihood estimates using the times of death. The d_1 deaths from cause 1 have been observed at times t_1, the d_2 deaths at times t_2.

It will be convenient, as before, first to estimate $\beta = \beta_1 + \beta_2$.

For a death at time t among the $d = d_1 + d_2$ deaths, the probability is $\beta e^{-\beta t}$, and for a survivor to the end of the period, the probability is $e^{-\beta}$. For the total sample the probability is proportional to

$$\phi = \beta^{d} e^{-\beta \Sigma t} e^{-\beta s} . \tag{30}$$

From this we derive the maximum likelihood estimate and its asymptotic variance [4], [5], [9]

$$\hat{\beta} = \frac{d}{\Sigma t + s} \tag{31}$$

$$\sigma^2_{\hat{\beta}} = \frac{\beta^2}{N (1 - e^{-\beta})} . \tag{32}$$

For the estimate of β_1 and β_2, we write the probability of the observations of the numbers and times of death from cause 1 and cause 2, and the survivors to the end of the period. Then the probability of the observations is proportional to

$$\phi = \beta_1^{d_1} e^{-\beta \Sigma t_1} \beta_2^{d_2} e^{-\beta \Sigma t_2} e^{-\beta s} \tag{33}$$

where $\beta = \beta_1 + \beta_2$.

From (33) we derive the maximum likelihood estimates and their asymptotic variances.

$$\hat{\beta}_1 = \frac{d_1}{\Sigma t + s} , \quad \hat{\beta} = \frac{d_2}{\Sigma t + s} \tag{34}$$

where $\Sigma t = \Sigma t_1 + \Sigma t_2$

$$(35) \qquad \sigma^2_{\hat{\beta}_1} = \frac{\beta_1^2 + \beta_1 \beta_2}{N (1 - e^{-\beta})}$$

$$(36) \qquad \sigma^2_{\hat{\beta}_2} = \frac{\beta_2^2 + \beta_1 \beta_2}{N (1 - e^{-\beta})}$$

COMPARISON OF THE FREQUENCY AND TIME ESTIMATES. Two sets of maximum likelihood estimates have been developed, one based on the observed frequencies of death from each cause, the other using also the times of these deaths. Presumably the time estimates, which use more "information," are better, and this should be reflected in a smaller variance of the time estimates. I shall compare the variances of the frequency and time estimates of β.

It is clear on inspection that the frequency estimate cannot be good for large β, $Q \to 1$,

$$\hat{\beta} = \ln (N/s) .$$

If Q is nearly unity the probability that s = 0, for even fairly large N, will not be small, and for all samples with s = 0, the frequency estimate of β is not determinable. In table 1 are shown the relative variances of the two estimates for different values of Q. It is seen that for small Q = .05 the variance of the time estimate is virtually equal to that of the frequency estimate. For $Q \leq 0.6$, the relative efficiency is greater than 0.9. Only with $Q > 0.9$ does the efficiency fall below 0.5. Since the frequency estimate requires only the number of deaths and not their times, and is easier to compute than the time estimate, it may be found satisfactory for use, except with very large Q.

MEASURE OF THE MORTAL EFFECT OF SMOKING. The ideas and formulas developed above are applicable to the analysis of the data of "prospective" studies into the relation of smoking and lung cancer. As a matter of fact Dr. Mindel Sheps [11], on the basis of a heuristic approach involving the notion of "exposed to risk," derived a maximum likelihood estimae which is identical with that developed here as the maximum likelihood frequency estimate of the net probability of death, from all causes, attributable to smoking. I shall consider the problem more in detail, in terms of the development I have outlined, particularly in respect of deaths from specific causes.

Consider deaths as segregated in two classes: those due to (1) some specific disease, for which I take lung cancer as an example, and (2) all other causes taken together. Non-smokers are subject to deaths from "natural causes." Smokers also are subject to death from natural causes, but we assume that, in addition, they are subject to deaths from lung cancer caused by specific carcinogens Y, and from other diseases caused

by substances X, these substances Y and X being contained in tobacco smoke. We assume that these causes act independently, and that the net probability of death, at time $< t$ $(0 \leq t \leq 1)$ in a unit period are given by

$$q'_{t_1} = 1 - e^{-\beta'_1 t} \tag{37}$$

$$q'_{t_2} = 1 - e^{-\beta'_2 t} \tag{38}$$

$$q_{t_1} = 1 - e^{-\beta_1 t} \tag{39}$$

$$q_{t_2} = 1 - e^{-\beta_2 t} \tag{40}$$

where q'_{t_1}, q'_{t_2} refer to net probabilities of death due to natural causes, from lung cancer and other diseases respectively, and q_{t_1}, q_{t_2} refer to death from lung cancer and from other diseases caused respectively by substances Y and substances X contained in tobacco smoke.

The corresponding observed crude probabilities of death are then

$$Q'_{t_1} = \frac{\beta'_1}{\beta'}\left(1 - e^{-\beta' t}\right) \tag{41}$$

$$Q'_{t_2} = \frac{\beta'_2}{\beta'}\left(1 - e^{-\beta' t}\right) \tag{42}$$

$$Q'_t = 1 - e^{-\beta' t} \tag{43}$$

$$Q_{t_1} = \frac{\beta_{11}}{\beta_T}\left(1 - e^{-\beta_T t}\right) \tag{44}$$

$$Q_{t_2} = \frac{\beta_{22}}{\beta_T}\left(1 - e^{-\beta_T t}\right) \tag{45}$$

$$Q_t = 1 - e^{-\beta_T t} \tag{46}$$

where

$$\beta' = \beta'_1 + \beta'_2$$

$$\beta_{11} = \beta'_1 + \beta_1 \; ; \beta_{22} = \beta'_2 + \beta_2$$

$$\beta_T = \beta_{11} + \beta_{22} = \beta'_1 + \beta_1 + \beta'_2 + \beta_2 .$$

N' nonsmokers have been observed, of whom d'_1 have died from lung cancer at times t'_1, and d'_2 have died from other diseases at times t'_2, while $s' = N' - d'_1 - d'_2$ have survived to the end of the period. We wish the maximum likelihood estimates of β'_1, β'_2, β_1, β_2, the corresponding net probabilities of death from lung cancer and from other diseases, attributable to natural causes, and attributable to cancer. We can derive these as before by writing out the probability of the total set of observations, including those on the nonsmokers and those on the smokers. However, the estimates may be had directly from the formulas already developed.

For the nonsmokers the parameters β'_1, β'_2, β', and the corresponding q's, and Q's which are functions of the β's are obtained directly from the formulas given, since these are the parameters involved in the exponential functions representing the probabilities of death among the nonsmokers. So far as the smokers are concerned, considering lung cancer, the deaths are due to (1) natural causes and (2) substances Y. We remember that in the exponential model the net risks are additive, so the exponential parameter of the smokers representing the risk for lung cancer is $\beta_{11} = \beta'_1 + \beta_1$. And similarly the exponential parameter representing the risk of death from other diseases among the smokers is $\beta_{22} = \beta'_2 + \beta_2$ as presented in (44), (45). Then β_{11} and β_{22} can be estimated from the observations on the smokers. Now, the observations on the nonsmokers and on the smokers are independent since they are made on different samples. So we obtain the estimate of β_1 and β_2 by subtraction

(47) $$\hat{\beta}_1 = \hat{\beta}_{11} - \hat{\beta}'_1$$

(48) $$\hat{\beta}_2 = \hat{\beta}_{22} - \hat{\beta}'_2 .$$

Similarly, since the estimates are independent, the variances are obtained as the sum of the variances of $\hat{\beta}'_1$ and $\hat{\beta}_{11}$. All the other estimates which are required are functions of the β's which have been estimated, and may be obtained by using the formulas for estimating functions.

The estimates of all the parameters involved in the analysis, with their variances, will be presented in a paper to be published later.* The chief parameter of interest here is the net probability of death due to a specified disease, here taken as lung cancer. I shall only write down the estimates for this parameter, which, it will be remembered, is symbolized by q_1.

(49) The frequency estimate is given most simply by,

$$\ln (1 - \hat{q}_1) = \frac{d_1}{d} \ln (s/N) - \frac{d'_1}{d'} \ln (s'/N').$$

The time estimate is given by

(50) $$\ln (1 - \hat{q}_1) = \frac{d'_1}{\Sigma t' + s'} - \frac{d_1}{\Sigma t + s}.$$

I take as an example of the application of the derived formulas, some data from the prospective study sponsored by the American Cancer Society and reported by Hammond and Horn [6], [7], [8]. Some 200,000 men in the age range 50 to 70 years were interviewed and a statement obtained from each as to his smoking habits. Periodically, inquiry was made and it was ascertained when any individual had died, and the time and cause of death as stated on the death certificate were recorded. A report was made based on the status of each individual as of 44 months after the initial inquiry. In table 2 are shown the essential data for the group of men 60-65 years of age at the time of the original inquiry. The binomial estimates of the probability of death in the 44 month period of follow-up are shown, for the nonsmokers and for the smokers, for each of four categories of cause of death, namely cancer of the lung, other cancer, coronary artery disease, and other diseases, as well as for death from all diseases. In the last 3 columns are shown three indices of the effect of smoking in increasing the probability of death from each of the categorized causes. The first of these indices is the estimate of the net probability of death from the respective causes, using the frequency maximum likelihood estimate. The second is the simple difference of the probabilities of death of smokers and of nonsmokers. The third column gives the so-called mortality ratio, which here is the ratio of the probability of death among smokers to the probability among nonsmokers.

If we use the net probability of death as the measure of the effect of smoking in respect of a cause of death, we see that, among the four categorized causes of death, smoking has the greatest effect in increasing deaths from coronary heart disease, the next greatest with diseases in the class "other diseases," the next from cancer other than lung cancer, and the least from lung cancer. If we use the simple difference of the probabilities of death, shown in the next column, we reach essentially

* Jointly with Dr. Lila Elveback.

the same conclusion. This is not surprising, since it is easy to show that if the probabilities of death Q, q are small, the net probabilities are given with close approximation by the simple difference of the probabilities of smokers and nonsmokers. If we use the "mortality ratio" shown in the last column, a quite different impression is obtained. We see that this index is 9.7 for lung cancer, while it is less than 2 for each of the other categories. In at least one important report from the United States Public Health Service [3], a ratio of less than 2 was considered as not even worth reporting as physically significant, and if this view is applied to the data of the table presented, it would only be said that these data show that smoking causes lung cancer. In an official statement on smoking by the Surgeon General [2] of the United States Public Health Service -- which depends largely upon the study represented by table 1, and other studies showing similar results -- only lung cancer is mentioned!

Which interpretation is valid -- that smoking is associated with death from all classes of disease, and chiefly from diseases other than lung cancer, or that drawn from the mortality ratio, which indicates a great effect on lung cancer and only a relatively small and negligible effect on any other disease?

The use of the mortality ratio has been criticized on a number of general grounds by Sheps [12] and by me [1] and it seems to the point to summarize some of them.

> If N animals are exposed to smoking and d_s die, while among N control animals d_o die, then the mortality ratio divides d_s by d_o to obtain a measure of the mortality due to smoking. In the conception of a death rate that places the number of "exposed to risk" in the denominator, this enumerates the dead controls as the "exposed to risk," and seems to imply that it is only those who are already dead from natural causes that can be killed by smoking! This has prompted Sheps ironically to title her article "Shall We Count the Living or the Dead?"
>
> It is arbitrary to use the ratio of mortality rates rather than the ratio of survival rates, and each gives a very different answer to the questions of the problem in hand.
>
> If a mortal drug were tried with controls, using the mortality ratio, it would appear to have a larger effect in a season when the natural mortality was low, than when it was tried in a season during which the natural mortality was high -- even if the actual effect of the drug was unaltered.
>
> Use of the ratio makes a small increase of deaths from a disease in the smoking group appear inordinately large if the natural mortality from that disease is small, and reduces to absurdity if the natural mortality is zero.

Yet there has been a great use of the mortality ratio in the studies referred to, with consequent emphasis on lung cancer. Now, the interpretation of the biologic significance of these statistical findings turns critically on how they are reported. If it is reported that smoking causes many diseases -- including such diseases as cancer of the prostate, for which no physical explanation is at hand -- it may be considered that the studies "prove too much," and that they are spurious, arising possibly from some unrepresentativeness in the sample. Or, if they are not spurious, they will perhaps be interpreted as reflecting a constitutional difference between nonsmokers and smokers rather than as supporting the theory that smoking causes lung cancer. But the general public, and also statisticians generally, have received the impression that all the statistical studies show is that smoking causes lung cancer. The public has not been told with anything like equal clarity, that smoking, in these statistical studies, seems to cause all classes of disease -- lung cancer only to the extent of about 15 per cent of the total. The statistical basis of this emphasis on lung cancer seems to be the use of mortality ratios, instead of net rates, or difference of death rates, to measure the putative mortal effect of smoking. This is the reason for linking the present paper on exponential competing risks to the statistical study of smoking and lung cancer.

Table 1

Comparison of Variance

Time estimate and frequency estimate

Q	N x Variance		Relative Efficiency
	Time Estimate	Frequency Estimate	
.001	.0010	.0010	1.000
.01	.0101	.0101	1.000
.05	.0526	.0526	1.000
.10	.1110	.1111	.999
.50	.9609	1.0000	.961
.60	1.3993	1.5000	.933
.90	5.8910	9.0000	.655
.95	8.9744	19.0000	.472
.99	21.4218	99.0000	.216

Deaths in 44 months. Age at beginning, 60-64 years

Cause of death	Nonsmokers* 20,278		Smokers** 21,594		Measure of effect of smoking		
	Deaths		Deaths		Net prob. due to smoking	Diff. of prob.	Mort. ratio
	No.	Prob.	No.	Prob.			
Lung cancer	10	.00049	103	.00477	.00449	.00428	9.7
Other cancer	218	.01075	325	.01505	.00469	.00430	1.4
Coronary disease	552	.02722	921	.04265	.01653	.01543	1.6
Other diseases	486	.02397	667	.03089	.00765	.00692	1.3
Total: All causes	1266	.06243	2016	.09336	.03303	.03093	1.5

* Nonsmokers = never smoked cigarettes regularly.

** Smokers = all regular cigarette smokers.

REFERENCES

1. Berkson, J., "Smoking and lung cancer: some observations on two recent reports," Journal of the American Statistical Association, 53 (1958) 28-38.

2. Burney, L. E., "Smoking and lung cancer," Journal of the American Medical Association, 171 (1959) 1829-1837.

3. Dorn, H., "Tobacco consumption and mortality from cancer and other diseases." Presented at the VIIth International Cancer Congress in London, July 8, 1958.

4. Epstein, B., and Sobel, M., "Life testing," Journal of the American Statistical Association, 48 (1953) 486-502.

5. Halperin, Max, "Maximum likelihood estimation in truncated samples," Annals of Mathematical Statistics, 23 (1952) 226-38.

6. Hammond, E. C., and Horn, D., "The relationship between human smoking habits and death rates," Journal of the American Medical Association, 155 (1954) 1316-1328.

7. Hammond, E. C., and Horn, D., "Smoking and death rates - report on forty-four months of follow-up of 187,783 men, I. total mortality," Journal of the American Medical Association, 166 (1958) 1159-1172.

8. Hammond, E. C., and Horn, D., "Smoking and death rates - report on forty-four months of follow-up on 187,783 men, II. death rates by cause," Journal of the American Medical Association, 166 (1958) 1294-1308.

9. Littell, A. S., "Estimation of the T-year survival rate from follow-up studies over a limited period of time," Human Biology, 24 (1952) 87-116.

10. Neyman, Jerzy, First course in probability and statistics, New York, Henry Holt and Company, (1950) See pp. 69-95.

11. Sheps, Mindel, C., "An examination of some methods of comparing several rates or proportions, " Biometrics, 15 (1959) 87-97.

12. Sheps, Mindel C., "Shall we count the living or the dead?," New England Journal of Medicine, 259 (1958) 1210-1214.

ARMY RESEARCH AND DEVELOPMENT

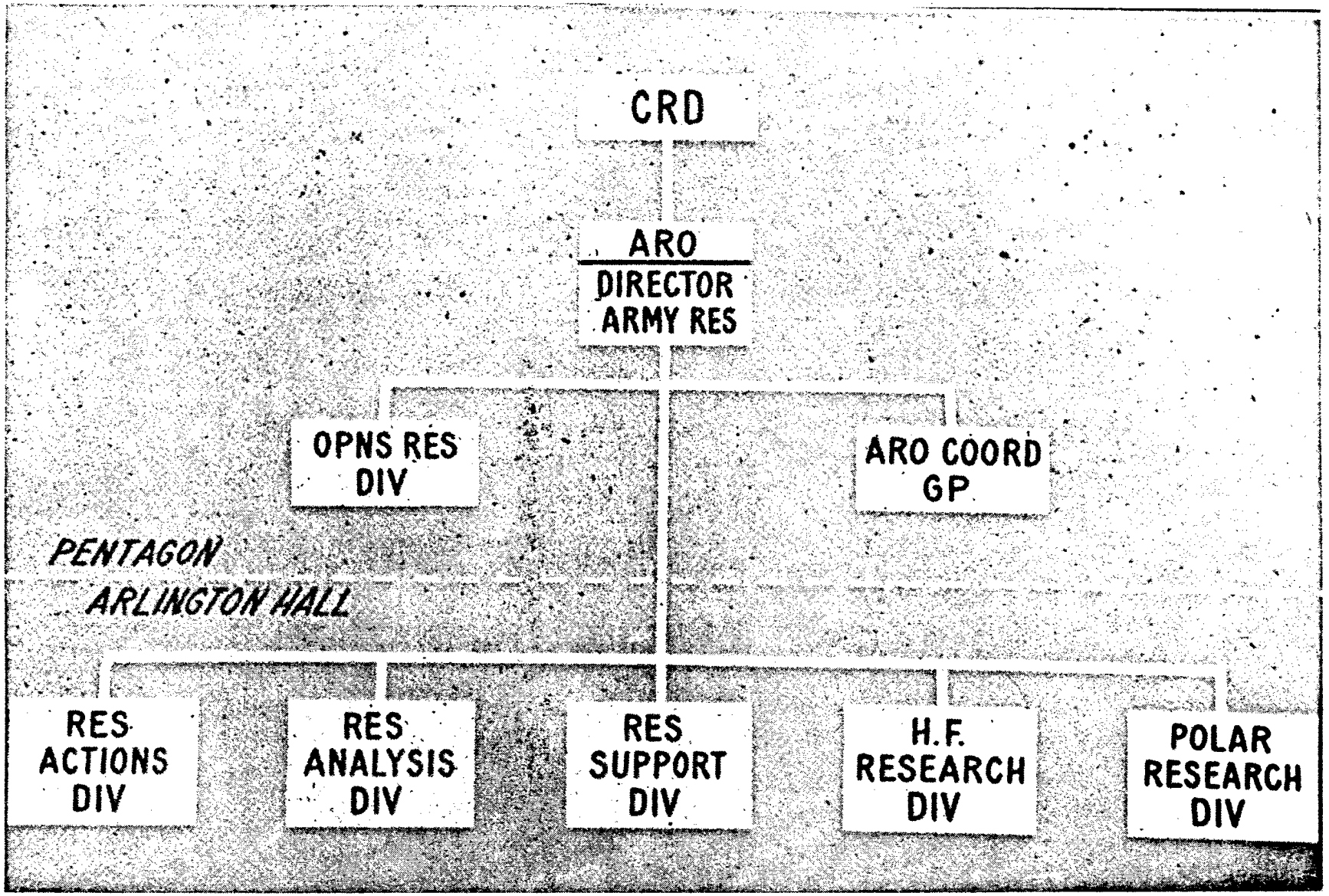
Richard A. Weiss
Army Research Office

Ladies and Gentlemen: What I want to talk about is Army Research and Development in its general aspects. Even though many of us are in the Army, it has been my experience that as members of one Technical Service, we tend to forget that there are other programs than our own. As a matter of fact, each Technical Service, in its own area, has responsibility for work of major importance in the research and development field. I thought if I could go through the program of the Army in a rather rapid fashion, it would give you some appreciation of the scope and possibly the depth of the programs being worked on by the seven Technical Services and their impact on the civilian economy.

We start with the organization of the Army. The organization of the Department of Defense might be likened to an onion, one of the many layers of which is the Army. The Chief of Research and Development is responsible for the planning and direction of research and development in the Army. He does this through three directorates: The Director of Developments, who is responsible for communications electronics, surveillance, the development of combat equipment, Army aviation, and the many developments necessary for the support of the ground soldier; the Director of Special Weapons, who is responsible for such areas as nuclear power, the nuclear aspects of missiles, and generally, with the overall weapons program of the Army; and the Director of Army Research, who has responsibility for monitoring the entire research program of the Army which is extensive and diverse. The Army Research Office, which is the operating element of the Directorate of Army Research, is, as you know, a rather newly formed office. It is in the process of being staffed and hopes to be in such a position as to present a sound scientific Army position to the outside scientific community and also do the job that it has to do to defend the support of basic research at the Defense level and in the Army Staff itself.

Slide 1

ARO Organization

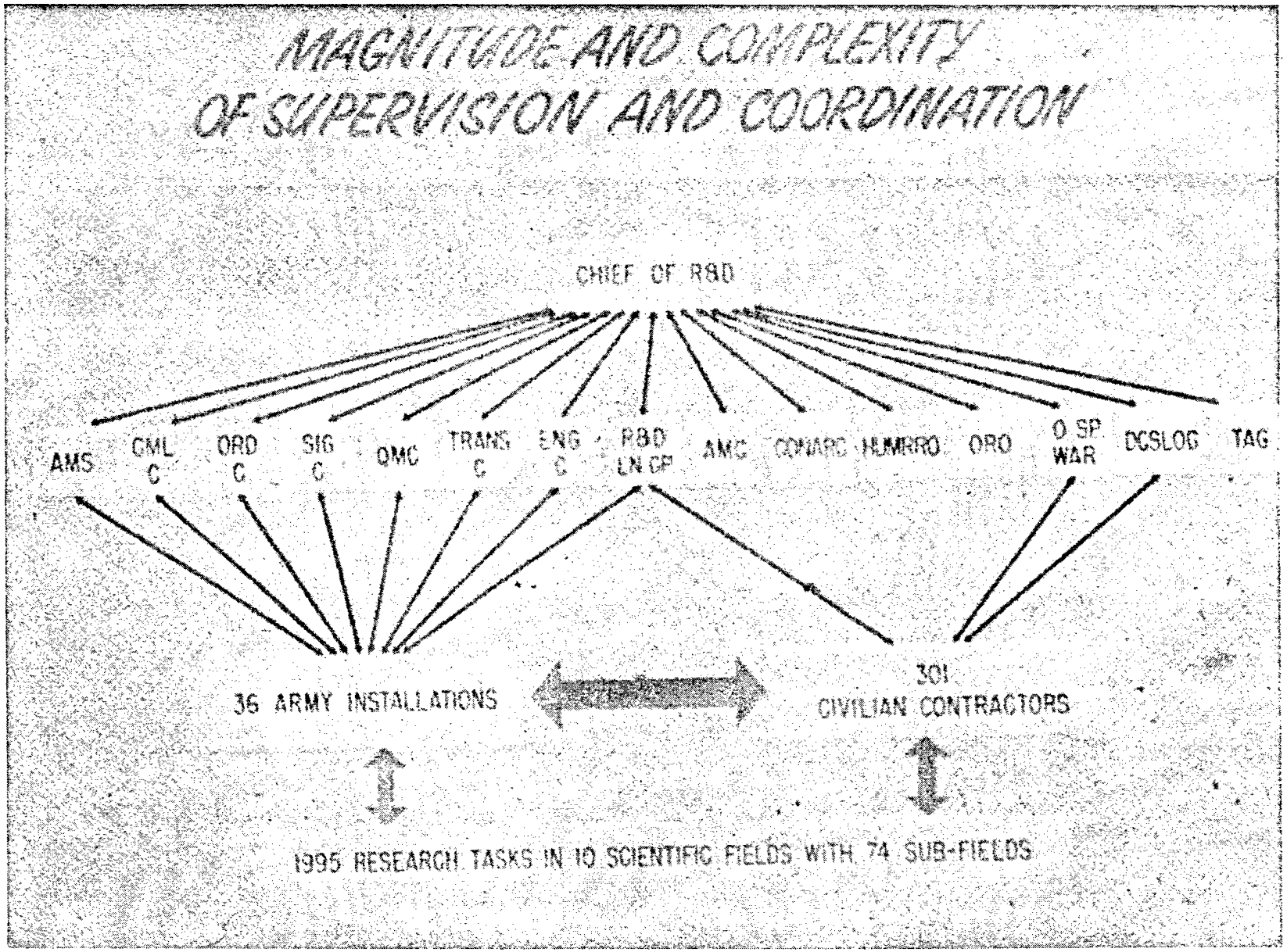


Slide 1

The first slide is a representation of the present Army Research Office organization. As you see, there is an Operations Research Division. This division is responsible for special studies cutting across much of the Army's overall mission and, particularly, those relating to research in the Army's operational problems. The Research Support Division has responsibitity for activities relating to scientific manpower, scientific information, and symposia and conferences of a scientific nature. It has just completed necessary staffing on a tri-service grants program. Human Factors Division is concerned with the problems of training and leadership and the relationship of the soldier to the machine. The three scientific divisions - Environmental Sciences Division, Life Sciences Division, and Physical Sciences Division - are composed of civilian and military scientists, all specialists in various scientific disciplines. In their particular fields, they analyze the program to determine gaps, determine the proper program balance, and develop policies effecting the improvement of the environment of the scientists in various laboratories and arsenals in the Army.

Slide 2

Magnitude and Complexity
of CRD



Slide 2

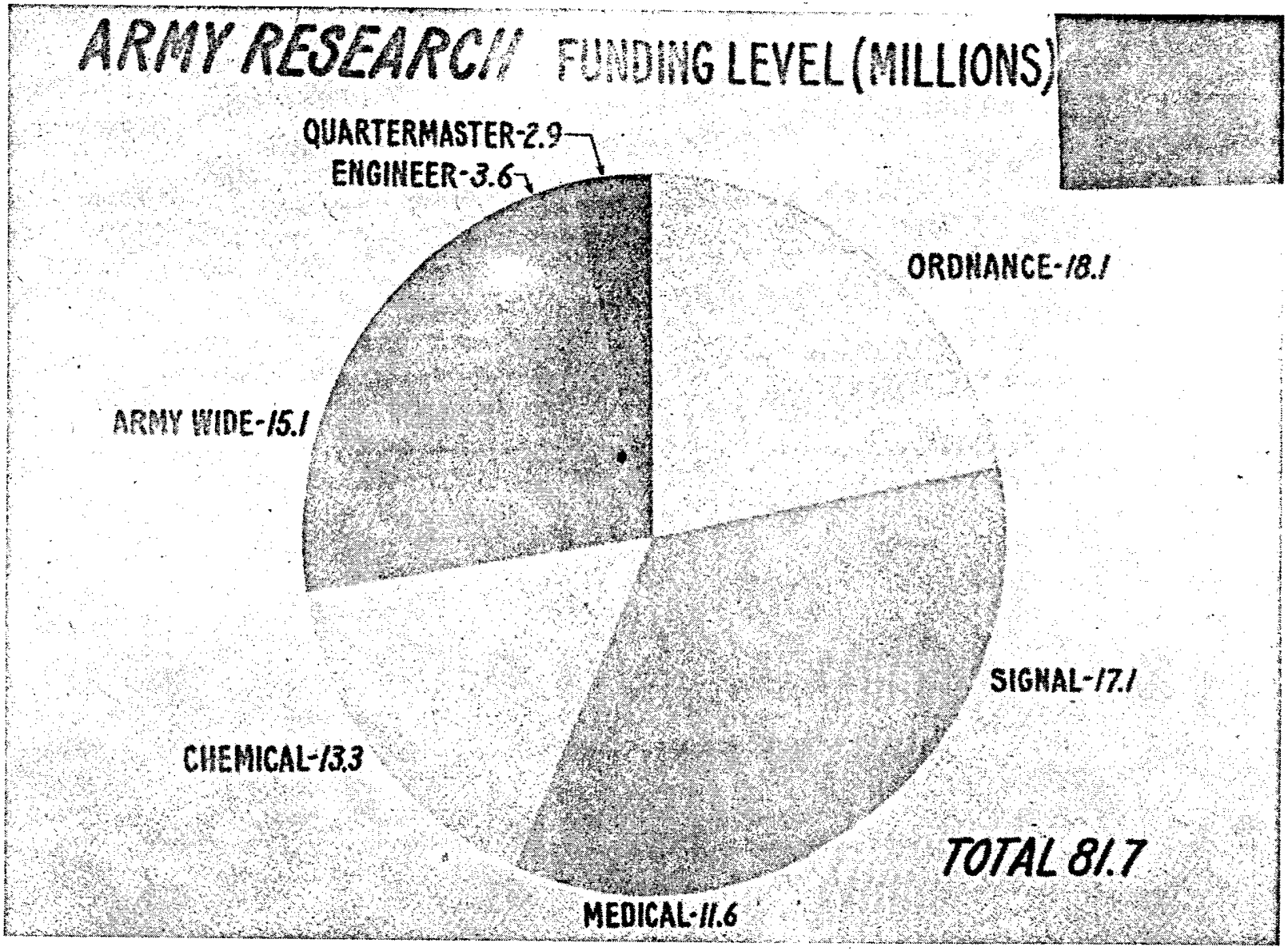
I don't know whether this can be seen, but this is an idea of the magnitude of supervision and coordination that the Chief of Research and Development has. The first seven blocks on the left are the seven Technical Services - Army Medical Service, Chemical Corps, Ordnance Corps, Signal Corps, Quartermaster Corps, Transportation Corps, and the Corps of Engineers. There is a Research and Development liaison group in Frankfort, Germany, carrying out the support of sciences in European communities. There is the Army Mathematics Center at the University of Wisconsin; R&D support for the Continental Command at Fort Monroe; The Human Factors Research Office at George Washington University; The Operations Research Office at Johns Hopkins University; R&D support for a division of Special Warfare; and R&D support of operations research for the Deputy Chief of Staff for Logistics.

These are the areas where the Chief of R&D has to provide funds for the support of the work that is going on. It is his responsibility to get the funds to carry out the mission and to provide support for the scientific staff. There are 38 Army installations where the research is being done. This is not generally known by most people in the Army. There are 19 government and over 400 civilian contractors, and as of today there are about 2400 research tasks in the various scientific disciplines covering something in the order of 74 sub-fields.

Now, a little bit about funding. The Chief of R&D has the problem of maintaining a balanced program, not only with the logistics of production but, once having got his share, maintaining a balanced program between the development program and the research program. And this is a rather trying task because each of these areas makes its own rather severe demands. Regarding the total funds in 1960: There is, roughly, a billion dollar budget - half a billion dollars in research and development and another half billion in test and evaluation. It will be divided, roughly, in the following fashion:

Slide 3

Funding (Pie View)

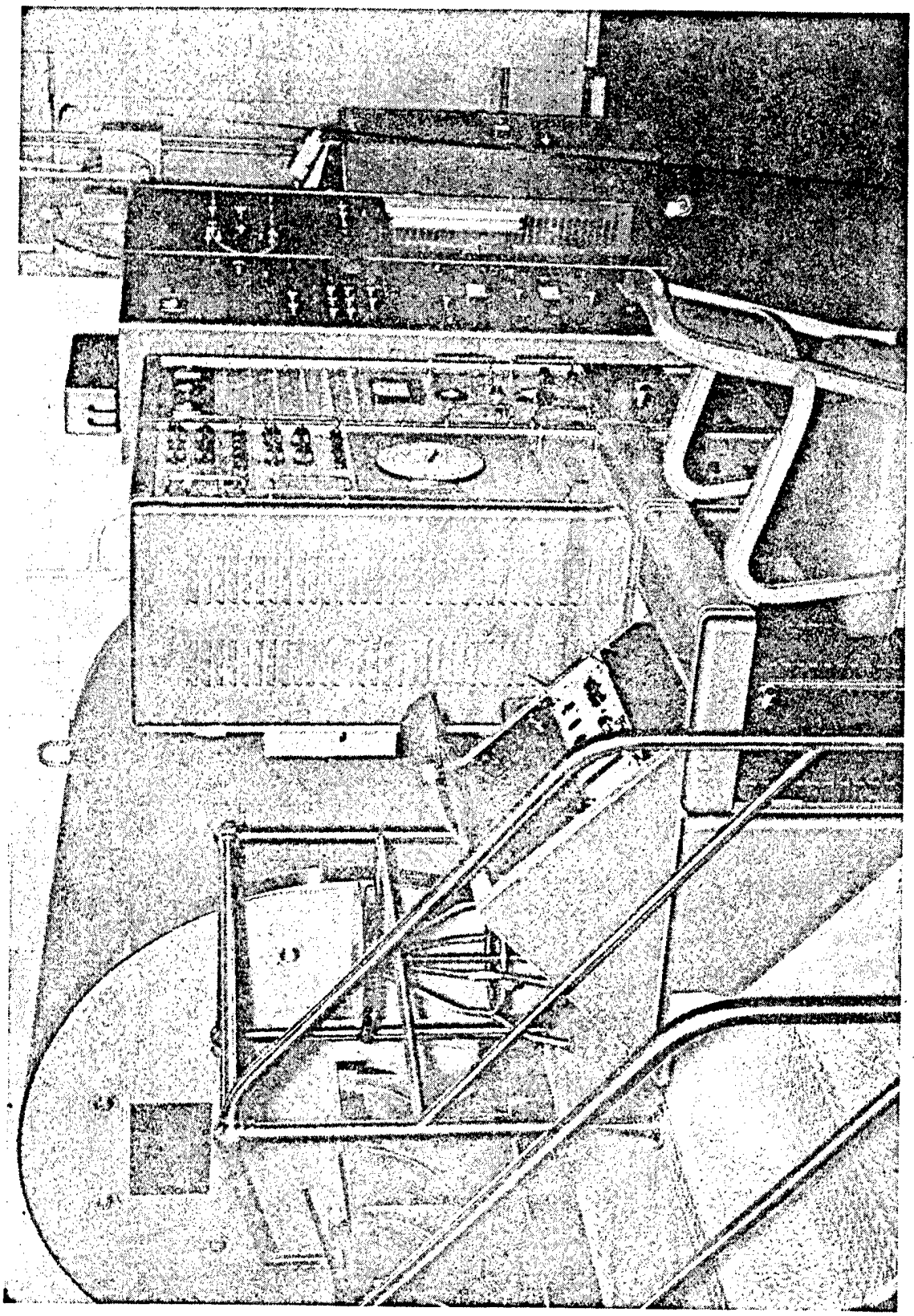


Slide 3

This slide shows a pie view of the allocation of funds among the seven Technical Services. You see Army-wide - $15.1 million - that has to do with a rather large program that goes across the entire board. Here the Office of Ordnance Research, the Army Mathematics Center, the European Research Office, and the new office in Japan are supported. Now, going around, you can see the relative proportions. This shows a total of $81.7 million in a prior year, but the program still balances up approximately the same, and the total research support, which is in the order of $130 million in 1960, will be divided among the Technical Services approximately according to the percentages shown here. Now, basic research is about 35 million dollars, which is of the order of one-fourth of the total Research and Development Program of $130 million. So I think that a ratio of one dollar for every four R&D dollars in support of research is a pretty healthy indication of the support of research that the Army is giving.

I would like to point out that the total contracts to non-government installations by all the seven Technical Services is in the order of 4,000. I think anyone would agree that in the research field alone, and I am not talking development or test and evaluation, the dollars that the Army spends certainly make a major impact on the total economy of the country. Obviously, the Navy and the Air Force do the same, the sum of money that is spent is considerable, and I would like to spend a little bit of time later indicating some of the things that have come out of the program.

Now, a few things about the Army installations.



Slide 4

Whole Body Radiation Counter

Slide 4

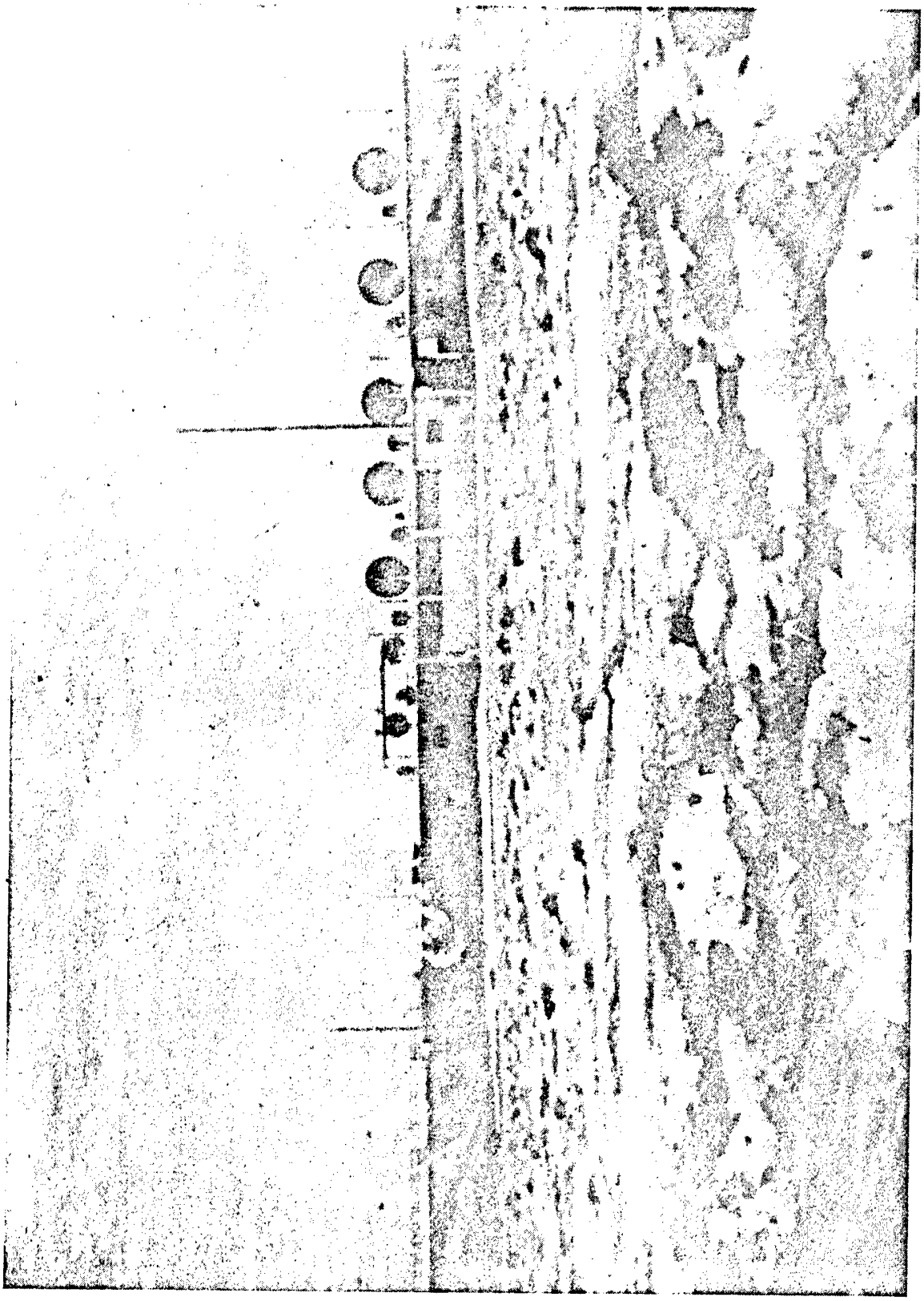
This is a slide of the whole body radiation counter at the Walter Reed Army Institute for Medical Research. It is possible to put a man inside so that the radiation emitted by virtue of any radioactive material that is in his body can be counted by a bank of scintillation counters. As a matter of fact, it is the only one, I believe, in the country, and much good research has been done at Walter Reed in this field.

Slide 5
(on following page)

Next is a slide of the White Sands Proving Ground which many of you will recognize. This is the Range Control Station for continuous plot of trajectories; a great amount of information is ground out here during the test flights.

Slide 6

The next slide shows an engine on a static test stand.

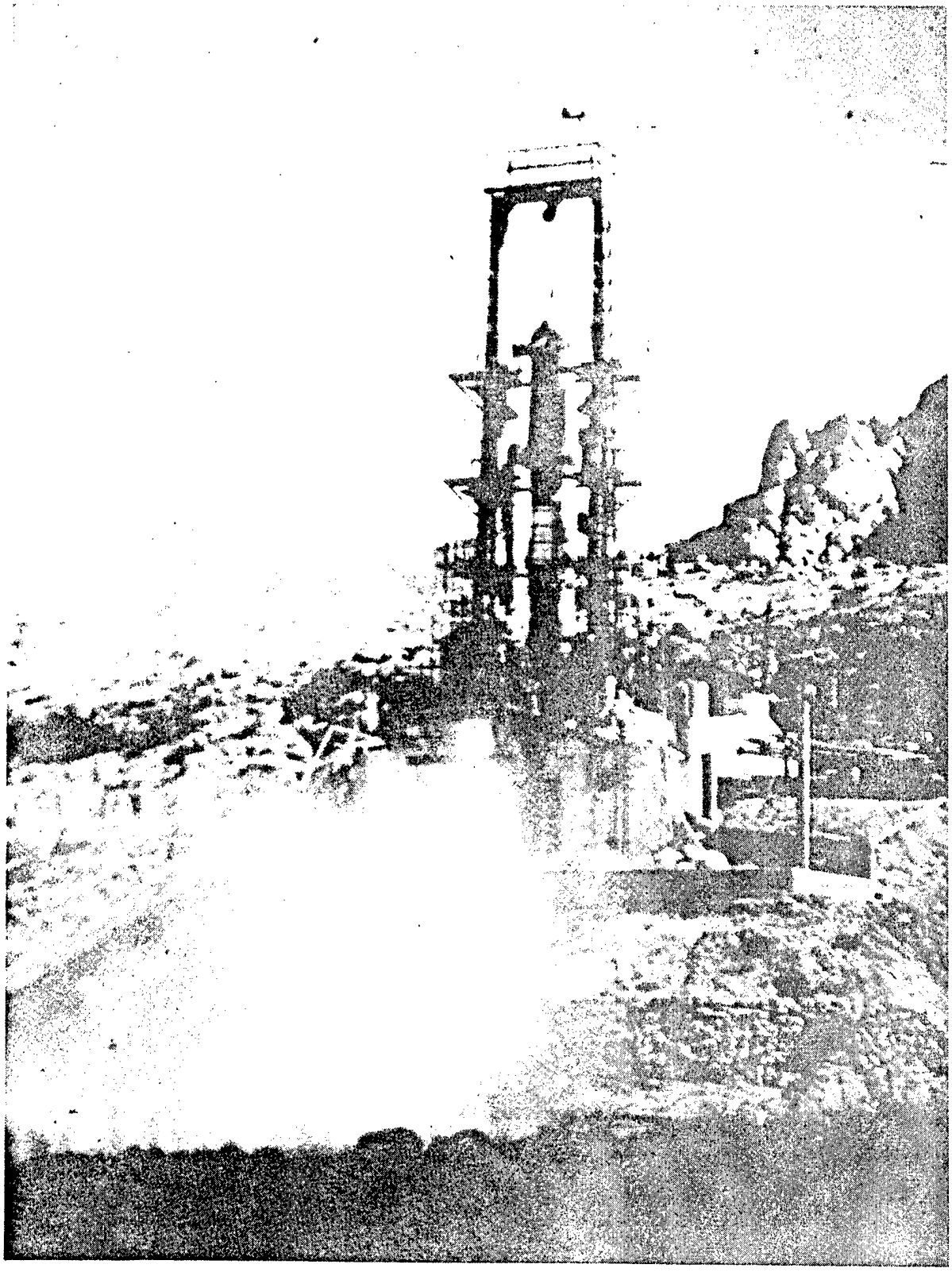


Slide 5

White Sands Proving Ground

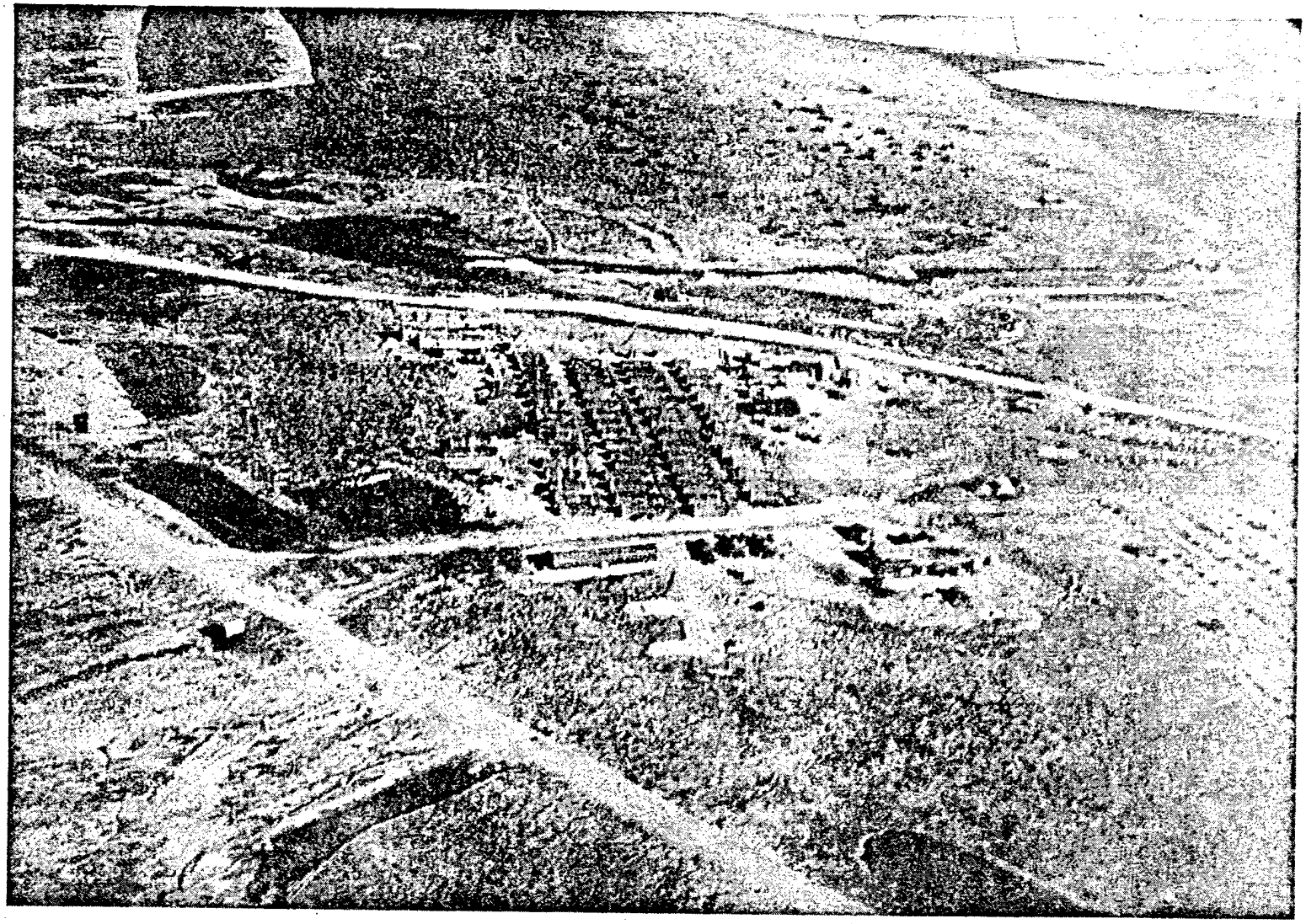
Slide 6

Engine on Static Test Stand



Slide 7

Thule, Greenland



Slide 7

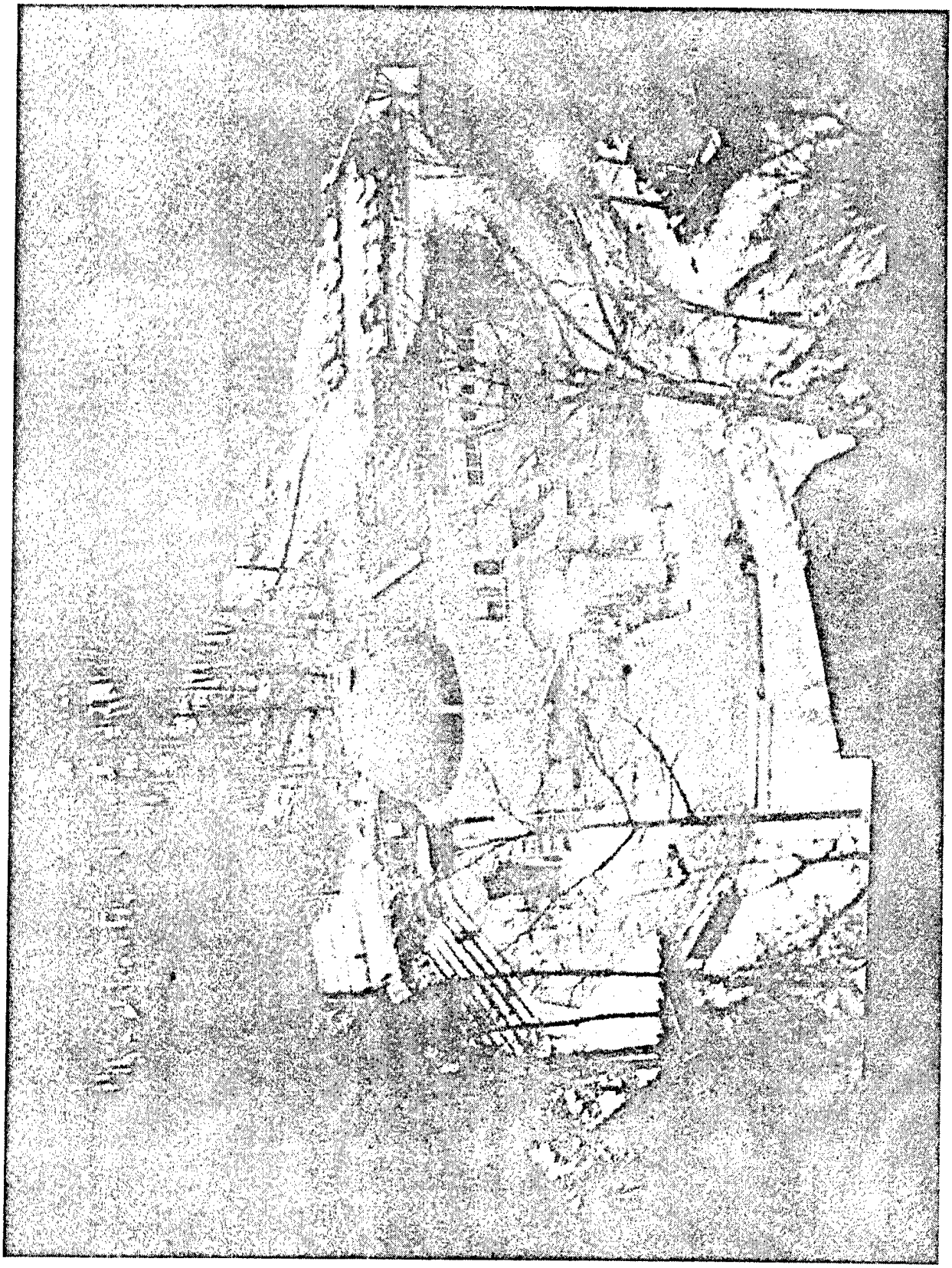
The following slide was taken in Thule, Greenland. Here there are something like nine disciplines covered in the 50 research tasks that are being undertaken by all the Technical Services. This is the area where the Engineers and the Transportation Corps, particularly, are engaged in a major program in determining how to survive and come to terms with the environment in these latitudes. Some rather remarkable and unique types of operations have been carried out by the Engineers under ice. Here ice is handled as one would cut stone from a quarry and ice construction is carried on underground. Laboratories have been established and experiments on ice and its characteristics are carried out.

Slide 8

This is something which will be recognized. It is the Army's packaged power reactor at Fort Belvoir, Virginia. The general feeling is that some rather remarkable things can be done in isolated areas with a facility of this kind. It is a 2-megowatt, thermally measured power apparatus, and is a prototype for others being built.

Slide 9

The next slide is of the solar furnace just recently installed at the Quartermaster Corps Natick Laboratory. The little white house in the center is the place where the beam is focused. There is a plane of mirror which tracks the sun and directs the parallel rays into a convex focusing mirror. Temperatures up to 3500 F. are achievable with fluxes of the order of 75 calories per sp. cm.

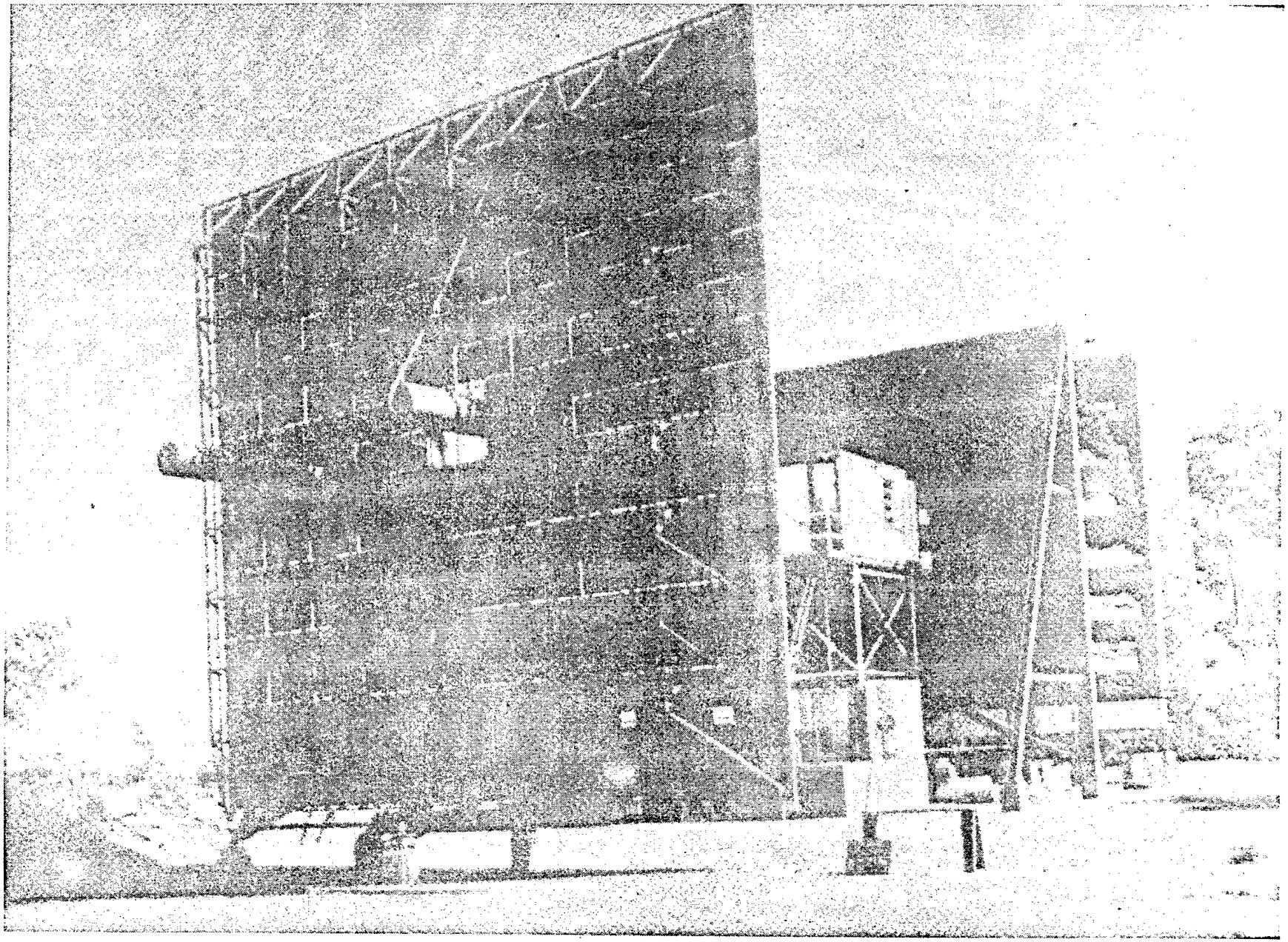


Slide 8

Power Reactor, Fort Belvoir

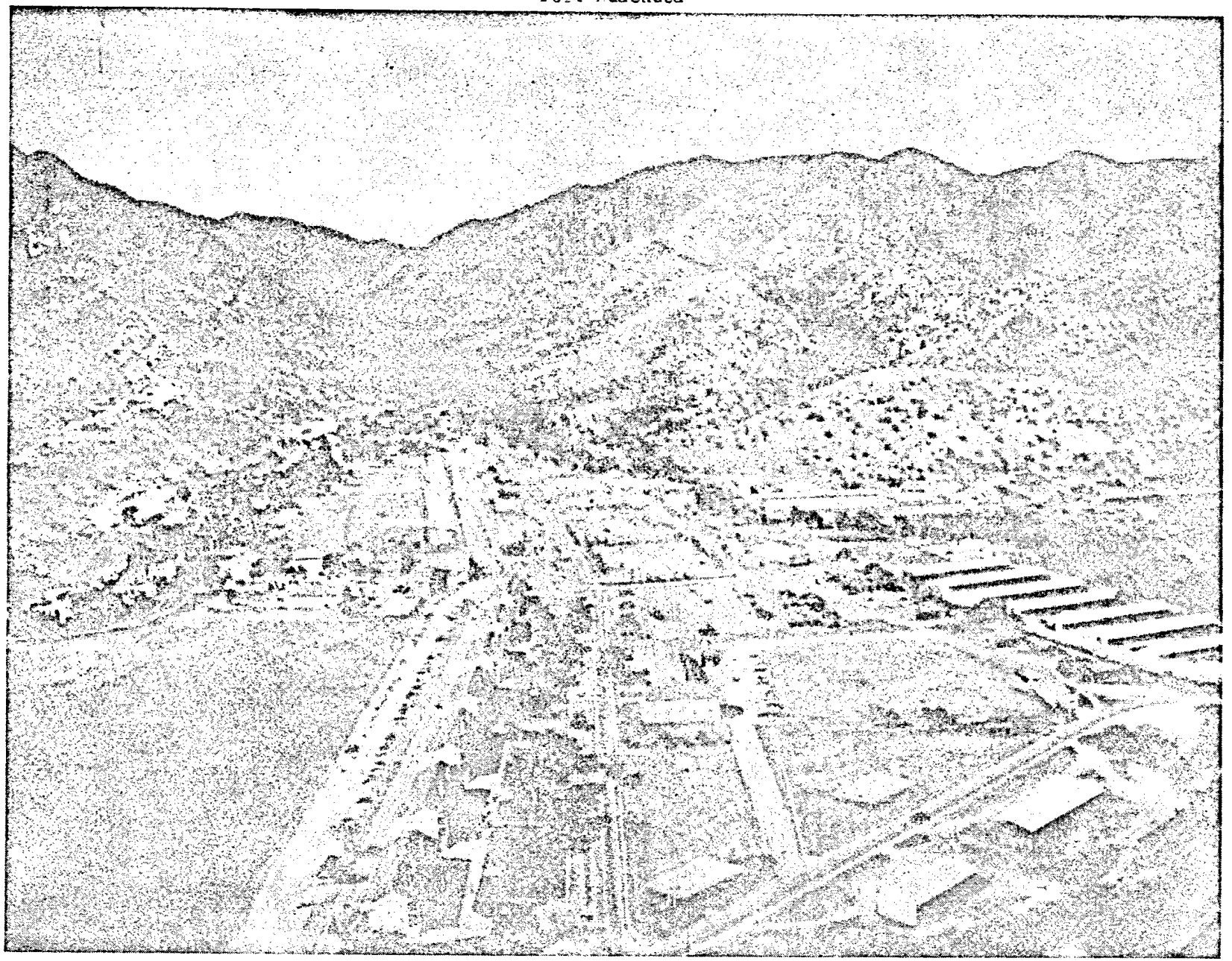
Slide 9

Solar Furnace



Slide 10

Fort Huachuca



Slide 10

This is a slide of a rather famous place. It is the old cavalry post at Fort Huachuca, in Arizona. It is now run by the Signal Corps and is an electronic proving ground where a great deal of testing is being done on surveillance devices, on the effect of counter-measures and general testing and evaluation in the field of communications for combat purposes. The red-roofed buildings are the old buildings that were there when the post was first established many years ago.

These are only a few of the number of installations that the Army has, and as a matter of fact, when one visits and sees the modern scientific and technical equipment available in the laboratories, one really has a great respect for the diversity and depth of Army science.

Now a few things about the accomplishments that the Army has been able to achieve which are of value to the civilian economy. Certainly World War II gave impetus to the aircraft industry and gave the chemical industry its greatest change to produce. It also ushered in the electronic age and the nuclear age. Following these, we now have the space age. Certainly one can expect, as time goes on, that the Department of Defense, with its three services, will be making other important contributions to our economy.

The Quartermaster Corps has done a major job in the processing and packaging of foods, much of which finds its way into the civilian economy. Pasteurization, dehydration, development of balanced diets for large groups, and minimum weight packaged material represent their area of contribution.

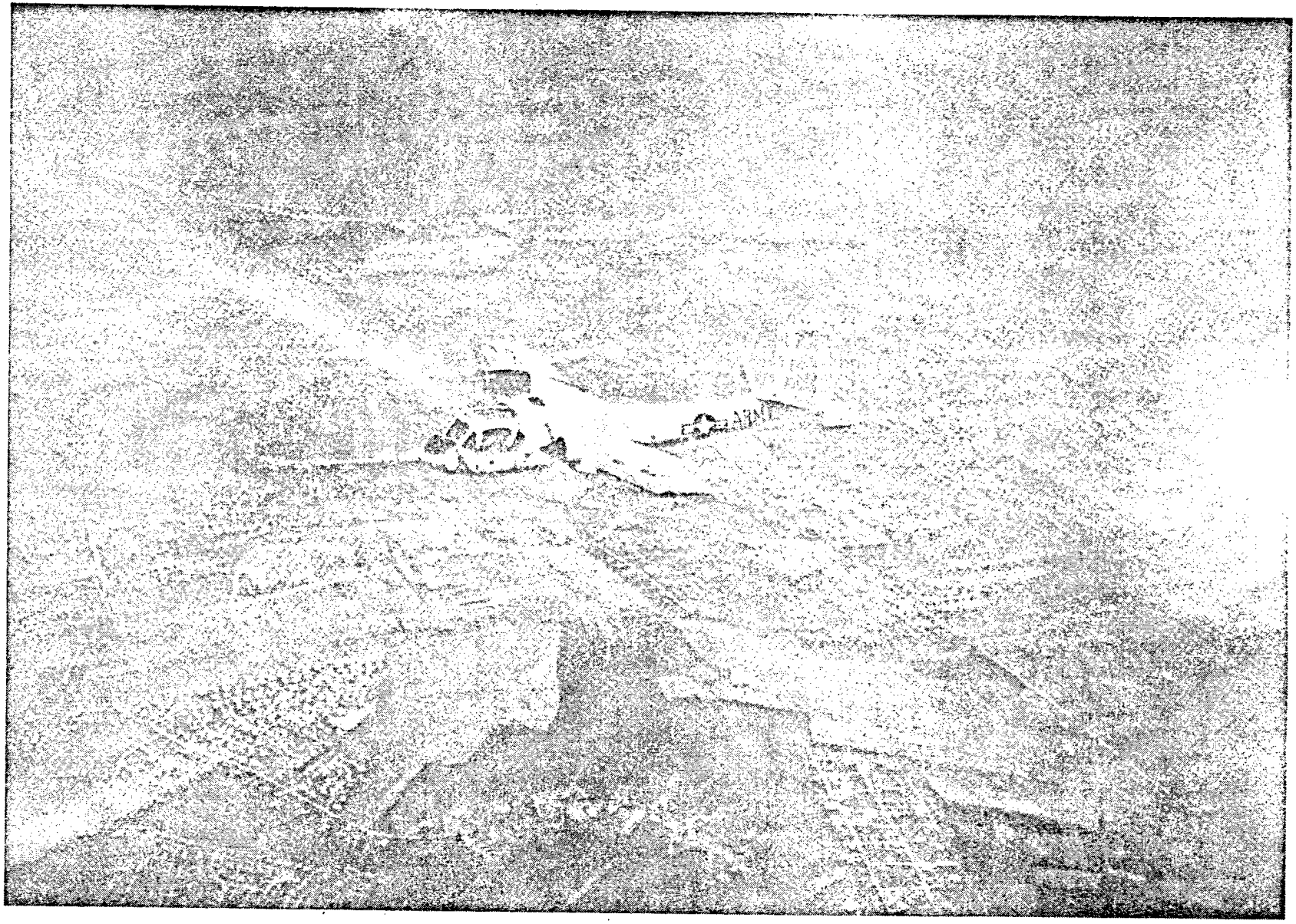
If the Communists ever use chemicals against us, we must be prepared to meet such an attack. Chemical Corps is working on this. On the other hand, Chemical Corps has performed recent tests which prove that nonlethal gases can incapacitate without killing, leaving no harmful after effects on humans or structures. Thus an objective can be captured without destroying needed buildings, bridges, or other man made structures. After receiving a dose of a gas of this type, humans will not react to orders or instructions, but wander around aimlessly. These gases are being investigated as possible alternatives to the massive exchange of thermonuclear weapons or the use of toxic agents.

In the Transportation Corps, they are concerned about advanced aircraft design, primarily of the type designated by "fly low, fly slow." The Army has to work close to the ground. Its vehicles have to be close to or on the ground, and much work is being done in this field.

I have a number of slides here showing some of the advanced designs - work that is in the development or prototype stage.

Slide 11

Mohawk



Slide 11

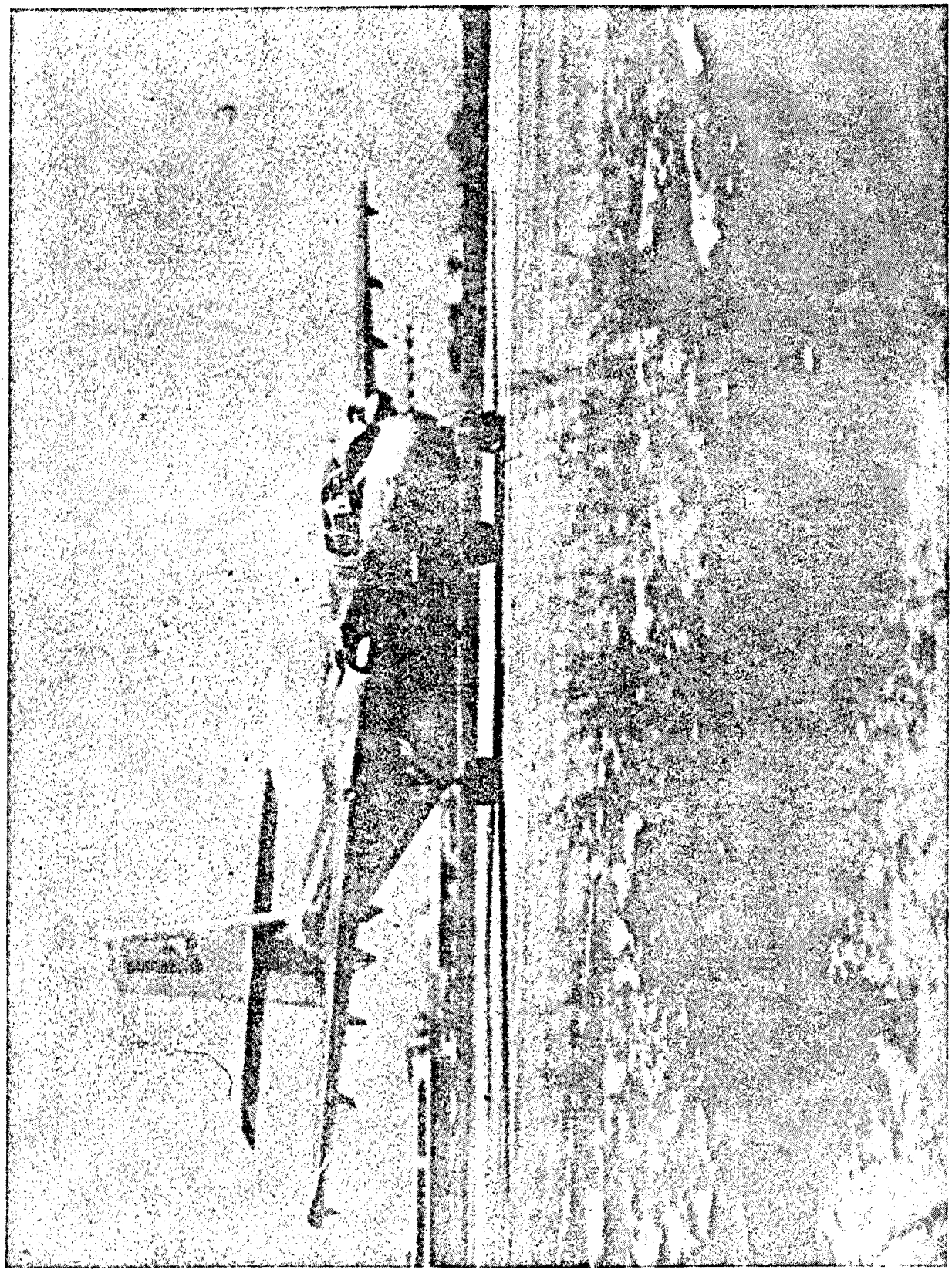
The first slide is the Army's Mohawk - reconnaissance and observation aircraft. This is a high-speed, short-take-off and landing aircraft for visual, photographic and electronic observations for shallow penetration into the enemy lines in the order of 25 to 40 miles.

Slide 12

The next slide is the Caribou. This is an airplane developed by the De Haviland Company in Canada. It is a three-ton short-take-off and landing aircraft, designed primarily for civilian use in Canada, Mexico and South America but now also serving the needs of the Army. As a matter of fact, this will be a plane which could be used wherever the development of a country has not advanced to the extent where you might expect to find prepared landing fields. This is particularly true in Brazil. I learned when I was in Brazil this last summer that Focke and a large staff of Germans left Germany shortly after the close of the War, went to Brazil and are now working for the Brazilian airforce in the development of VTOL and STOL aircraft.

Slide 13

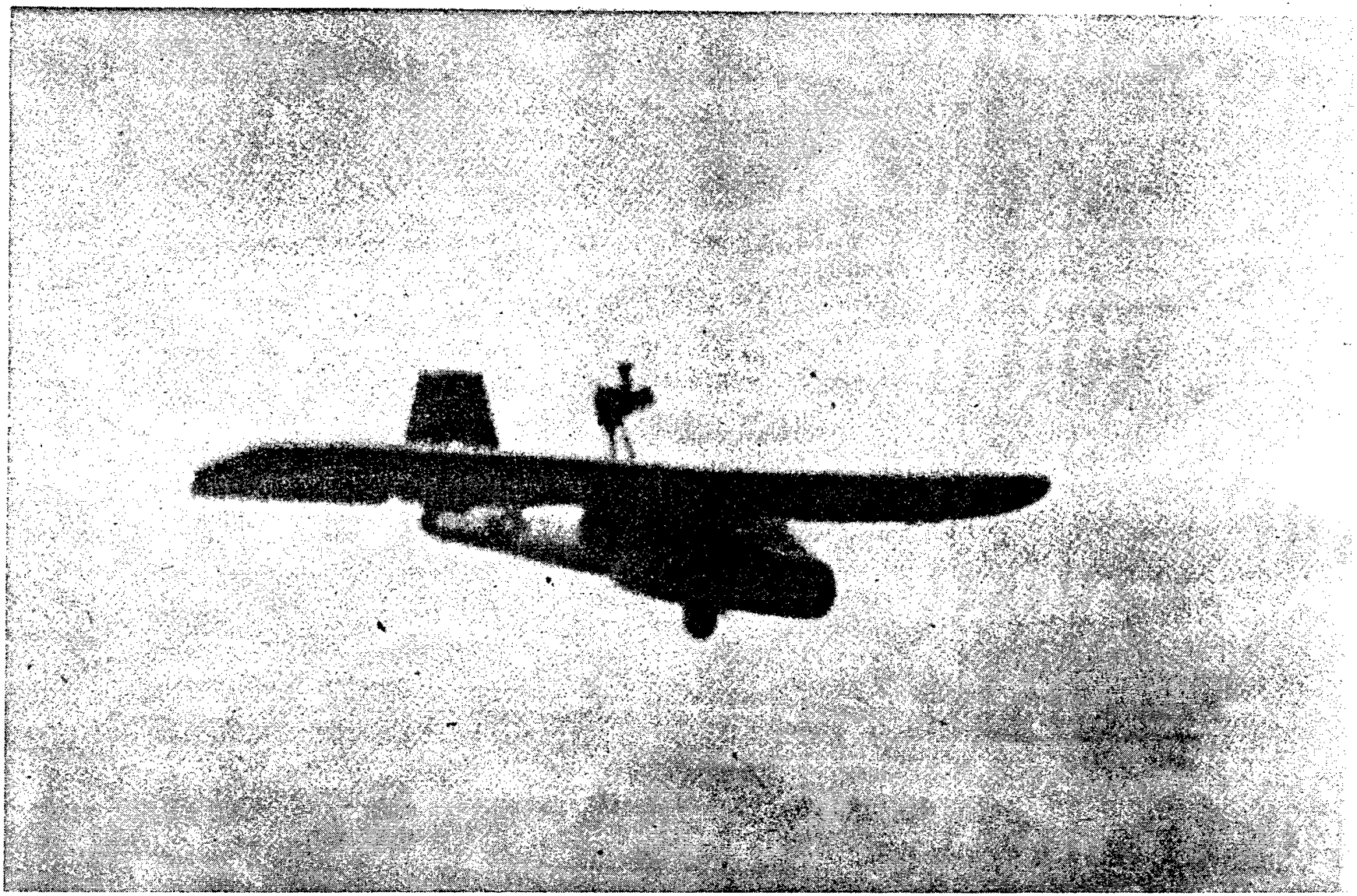
The next slide indicates work done by Goodyear on the Inflatoplane. This is a compact, rubberized craft which can be inflated by means of high-pressure CO_2. Parachuted out of aircraft, it can be assembled on the ground and flown from crude landing fields. Here it is, assembled and in the process of take-off. The engine, if you can see it, is right above the wing, a little to the right of the tail.

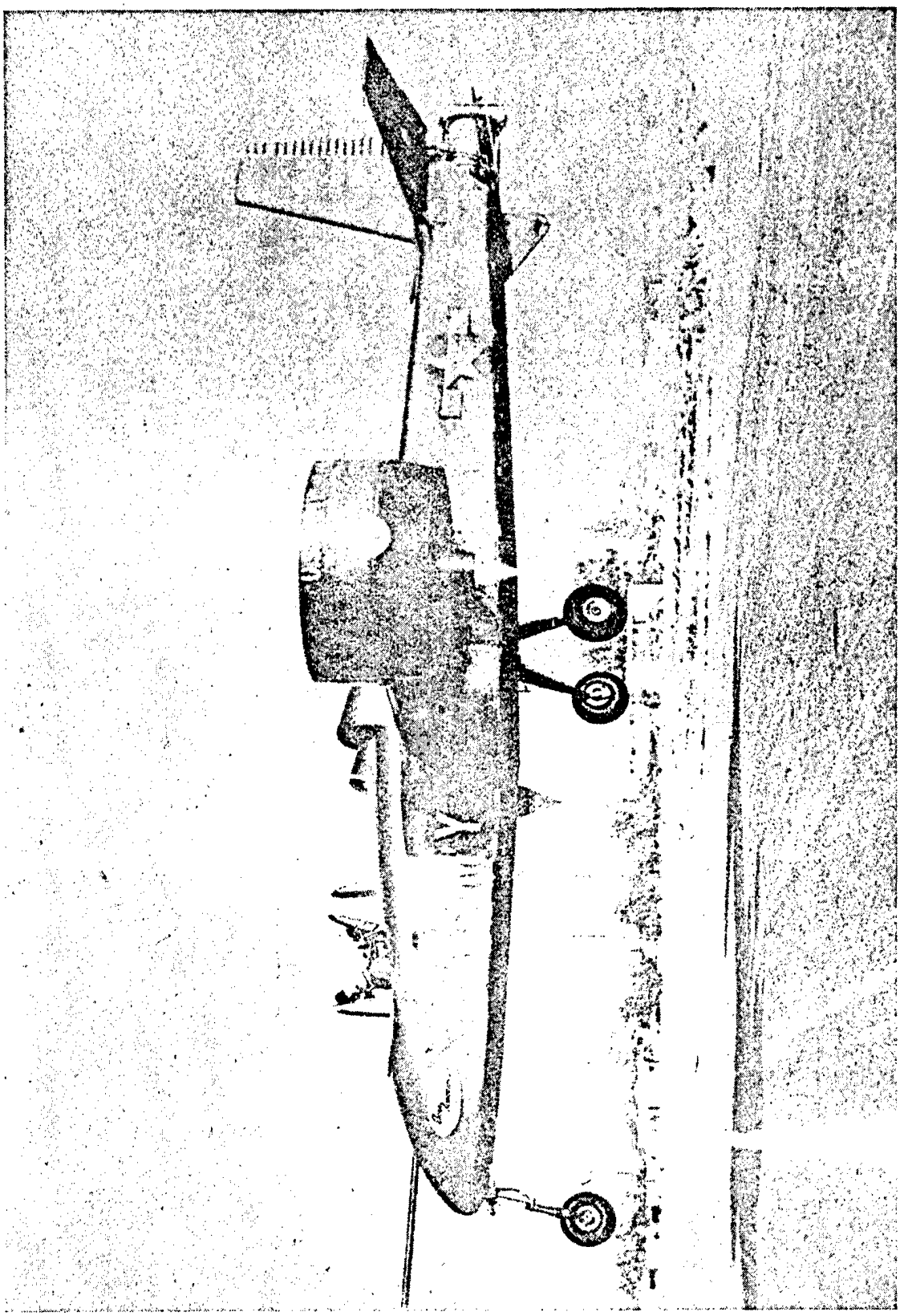


Slide 12
Caribou

Slide 13

Inflataplane





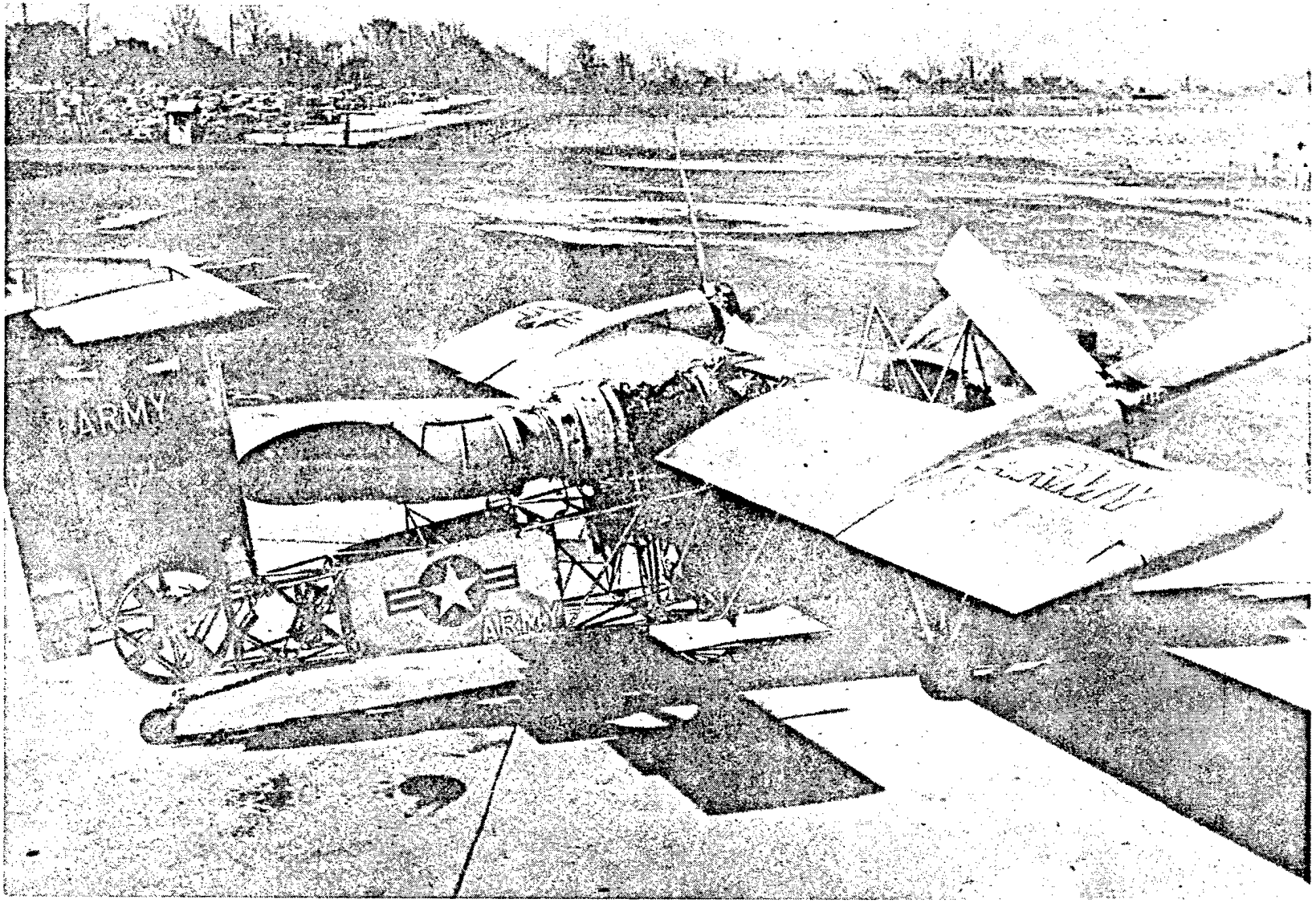
Slide 14
Ducted Fan Doak

Slide 14

A little bit about the air-column supported vehicles. These are research aircraft. When I talk about air-supported, I mean the ducted fan and convertor type planes. They are of the vertical take-off and landing (VTOL) and the short take-off and landing (STOL) type aircraft. The first one is a rotatable ducted fan made by Doak. Here you can see that the fan moves through its transition phase and lift phase and then it moves forward. The problems of transition, I understand, are difficult.

Slide 15

Vertol Tilt Wing



Slide 15

The next slide is of the Vertol tilt-wing aircraft; here the whole wing tilts instead of just the fan nacelle.

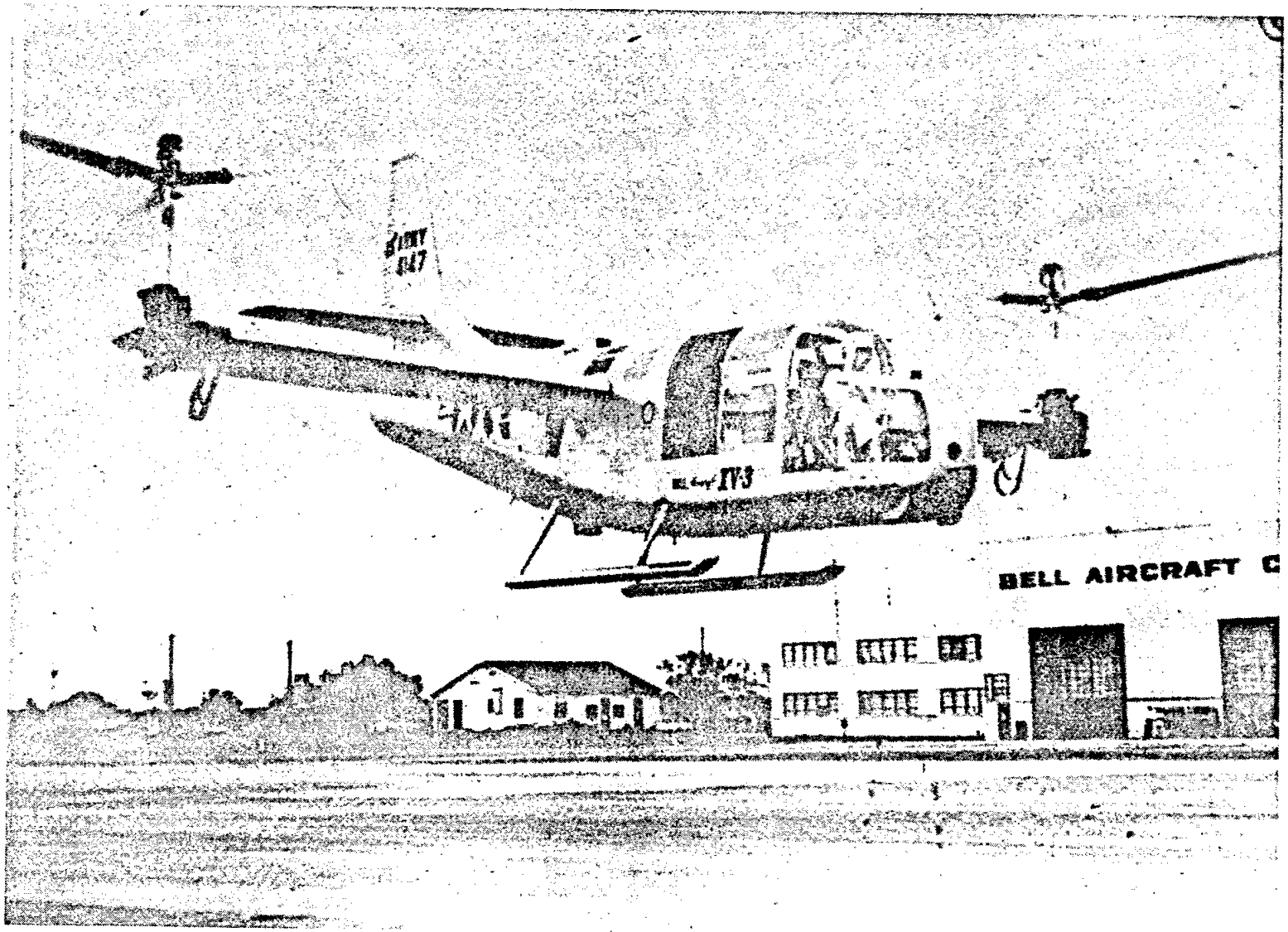
Slides 16, 17

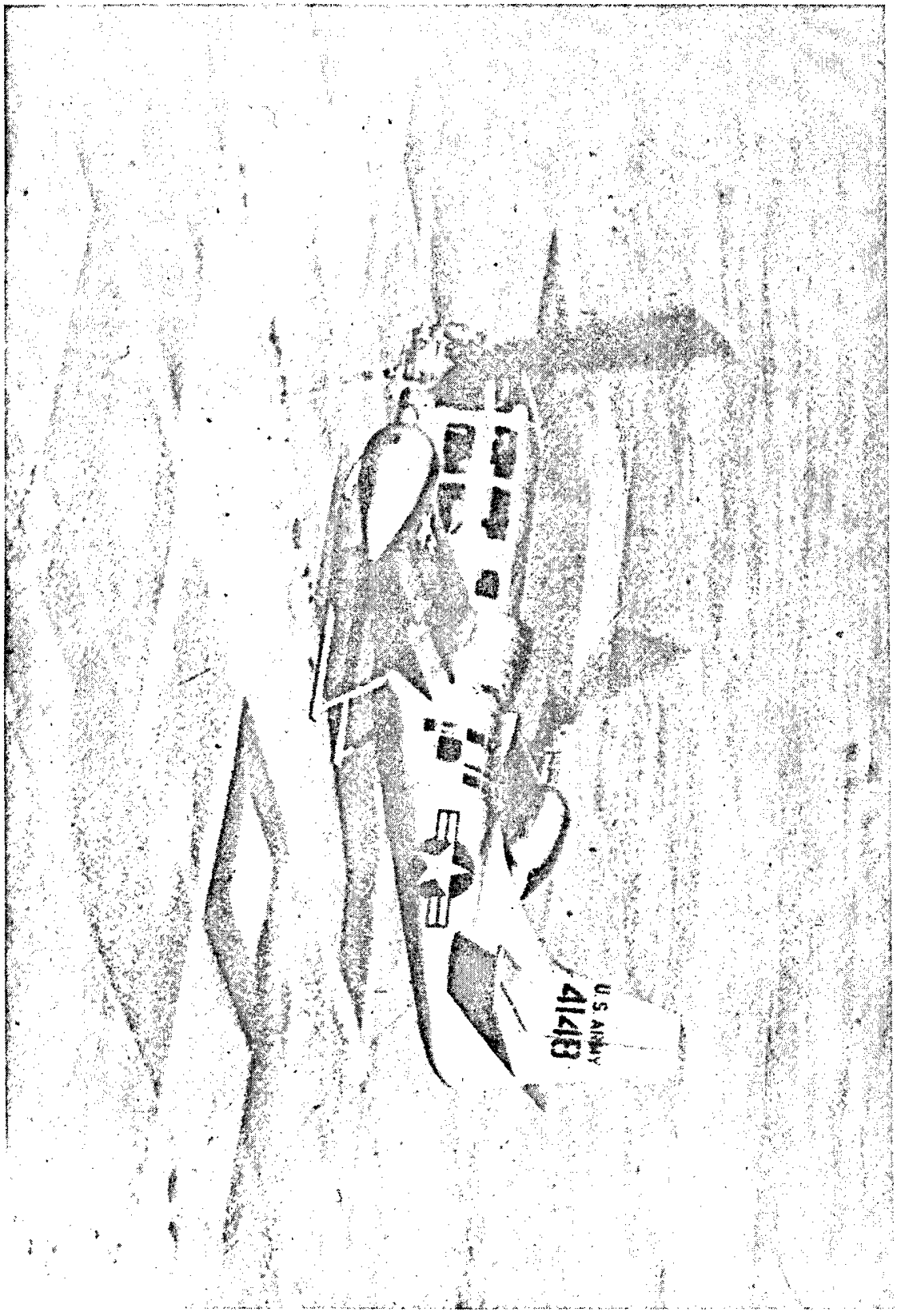
In the converti-plane made by Bell Aircraft you can see, in the next slide, the lift phase, and in the next, the forward flight phase where the propeller has turned through 90 degrees. They have successfully gone through the transition phases of several of these experimental aircraft.

Now, something about aerial vehicles. These are general purpose vehicles. There are a number of slides I am going to show of various designs.

Slide 16

Bell Convertiplane - Lift Phase

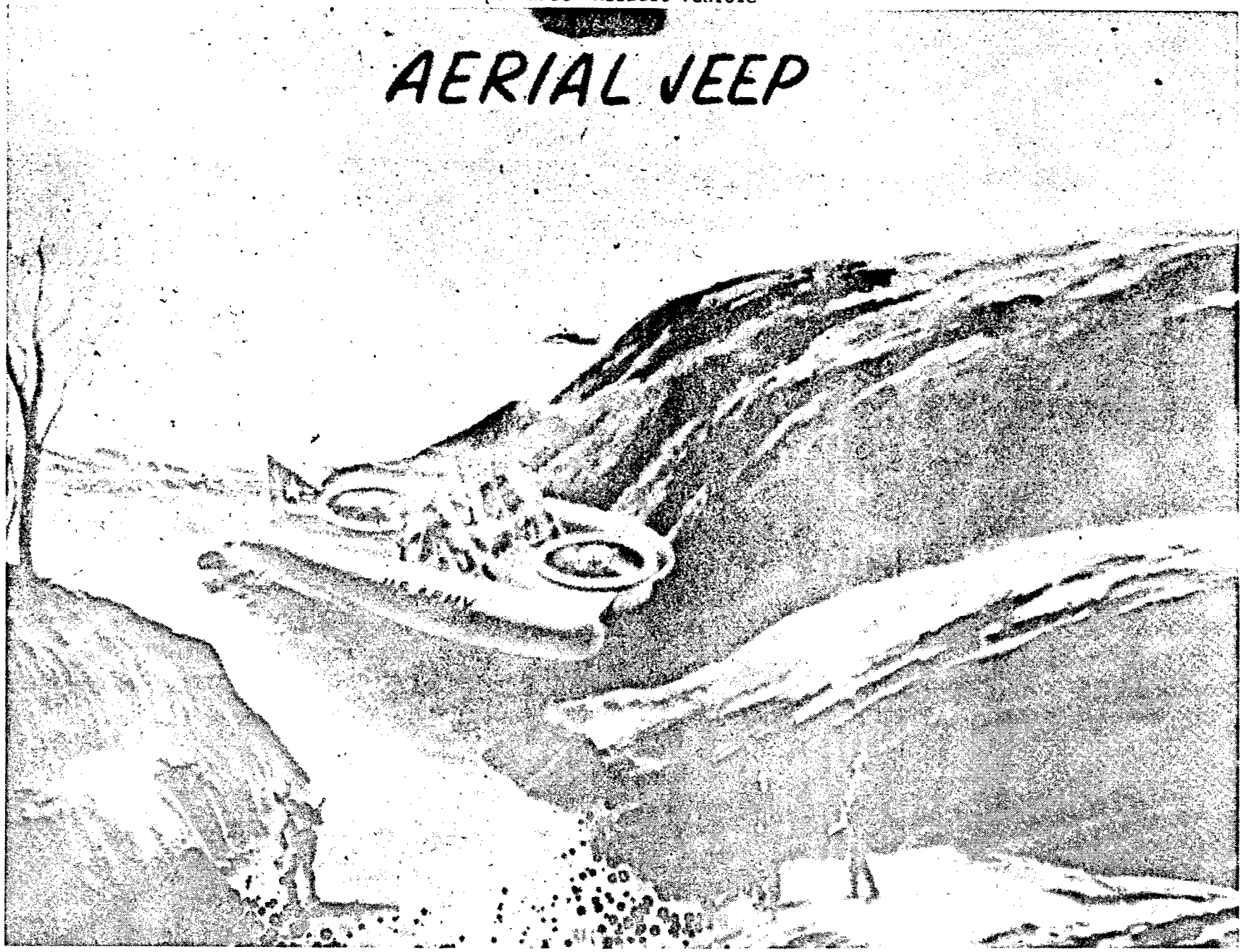




Slide 17

Bell Convertiplane - Forward Phase

Concept Phase Assault Vehicle



Slide 18

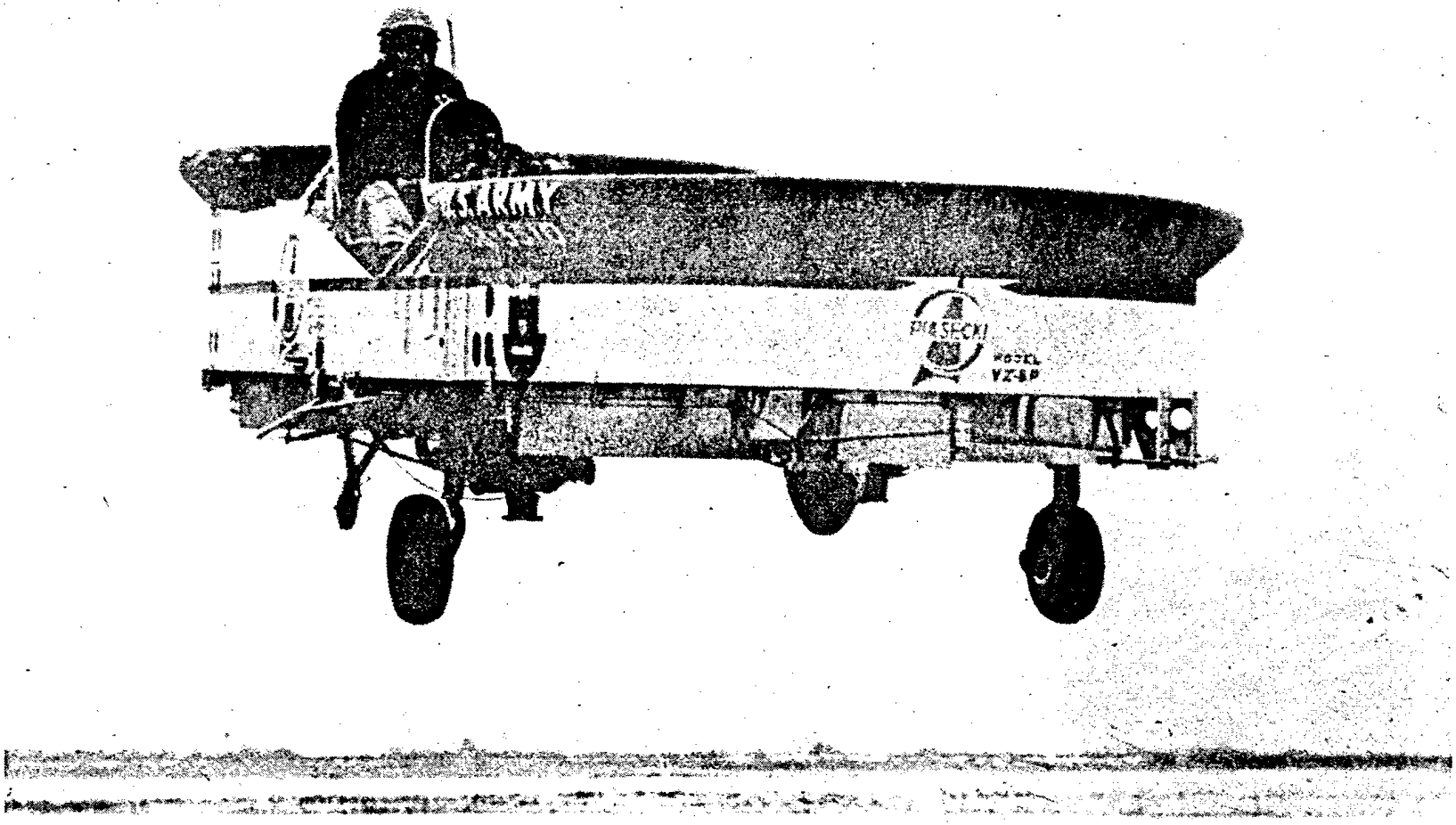
First is a Piasecki prototype aerial vehicle we used to call the "Jeep," however that is a patented name now, so we have to call it something else. It's just a flying vehicle - an artist's concept, as a matter of fact.

Slide 19, 20

The next slide is the developmental aircraft in flight and in the process of take-off, and the following slide is the same aircraft - that is Mr. Piasecki flying it and coming down to a landing. Someone has said that this vehicle has the glide angle of a rock, and I would guess that it probably has, if the engine failed.

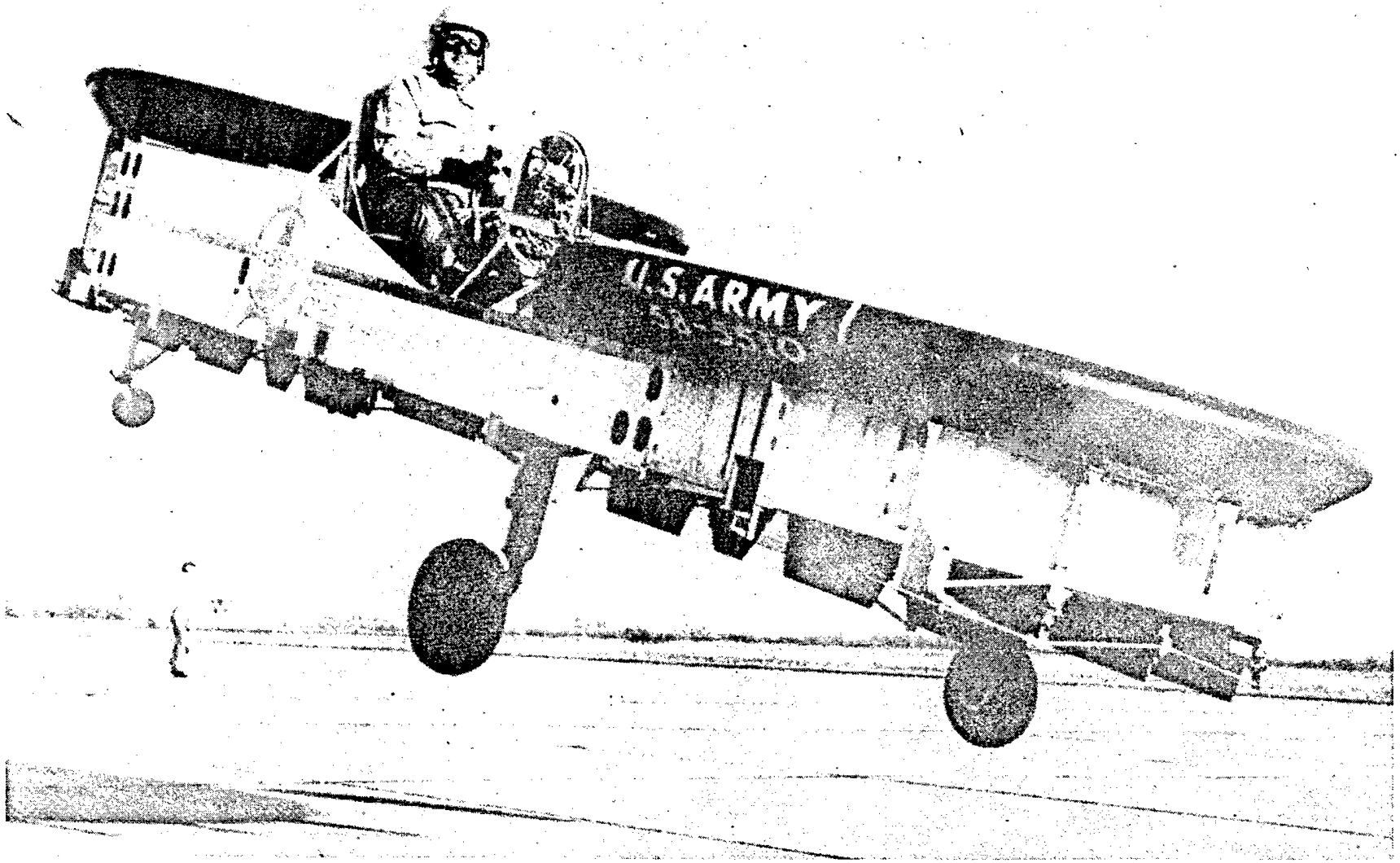
Slide 19

Aerial Jeep - Flight



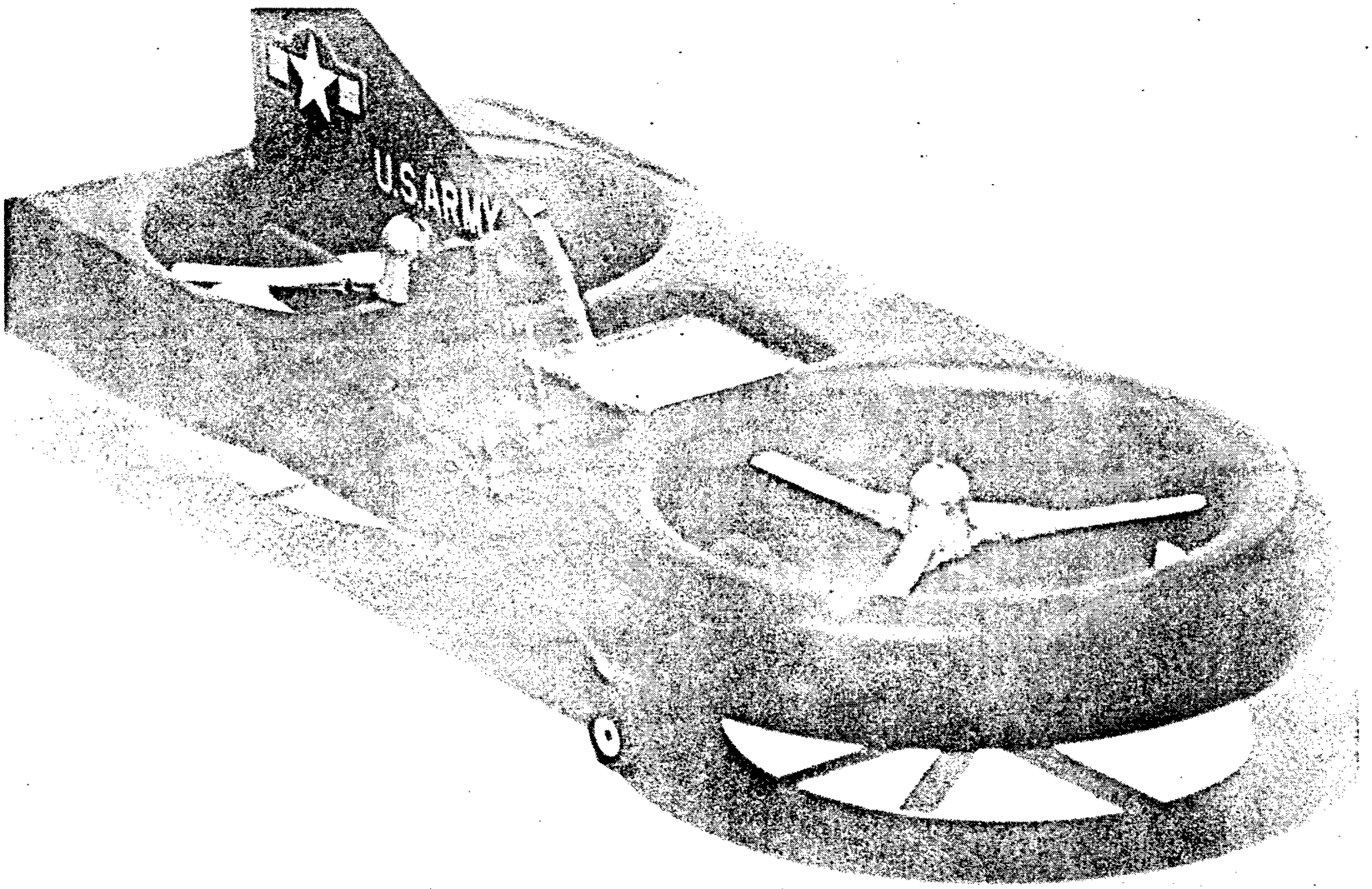
Slide 20

Piasecki - Aerial Jeep - Landing



Slide 21

Aerial "Jeep"



Slide 21

The next slide is a concept of the operational phase of an aerial jeep showing two down draft propelling systems and some idea of how it might make use of terrain cover.

I must mention here that the fly-low-fly-slow type of philosophy is governing the design and development of Army aircraft. It is designed to hug close to the earth and make use of all natural terrain so that problems having to do with observation can be carried out without danger of hazard from the enemy.

You know that the Army, and in particular the Ordnance Corps, has a major program in rockets. I thought you would like to get a look at three of them.

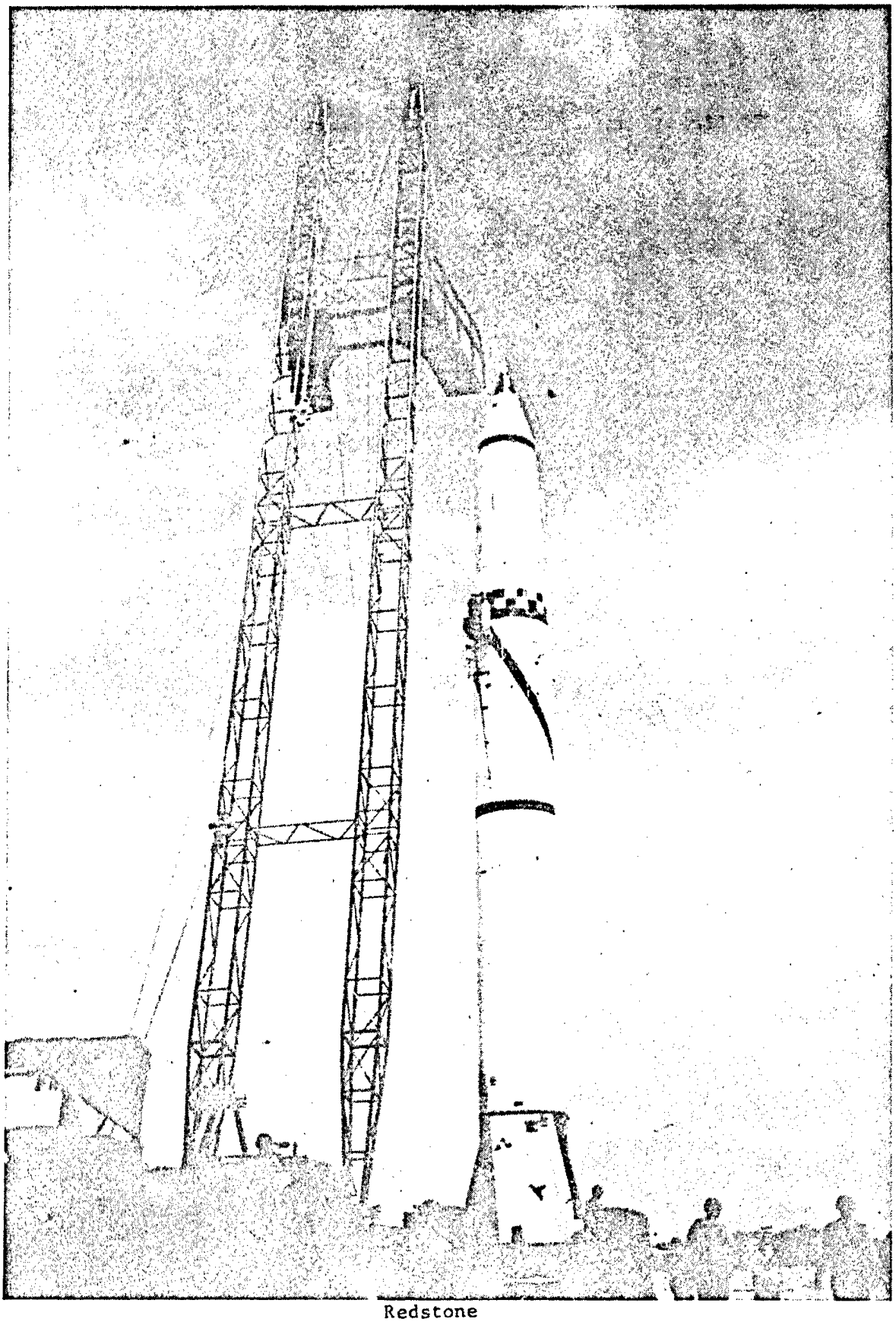
Slide 22

The first one is the Redstone, which, as you know, has proven its reliability. I do have a statement here which is the only one that the Army has been permitted to make in regard to space. "This missile (the Redstone) because of its proven reliability and stability, will be used to launch the first American into space as part of the NASA's Project Mercury." That is the extent to which I can say anything about it.

Slide 23

This is Jupiter C, or Explorer I, as you know it. It is a modified Redstone. It was the Free World's first satellite, and is expected to be up four or five years more. Judging from the number of successful firings it looks as though successful satellites are getting to be old hat.

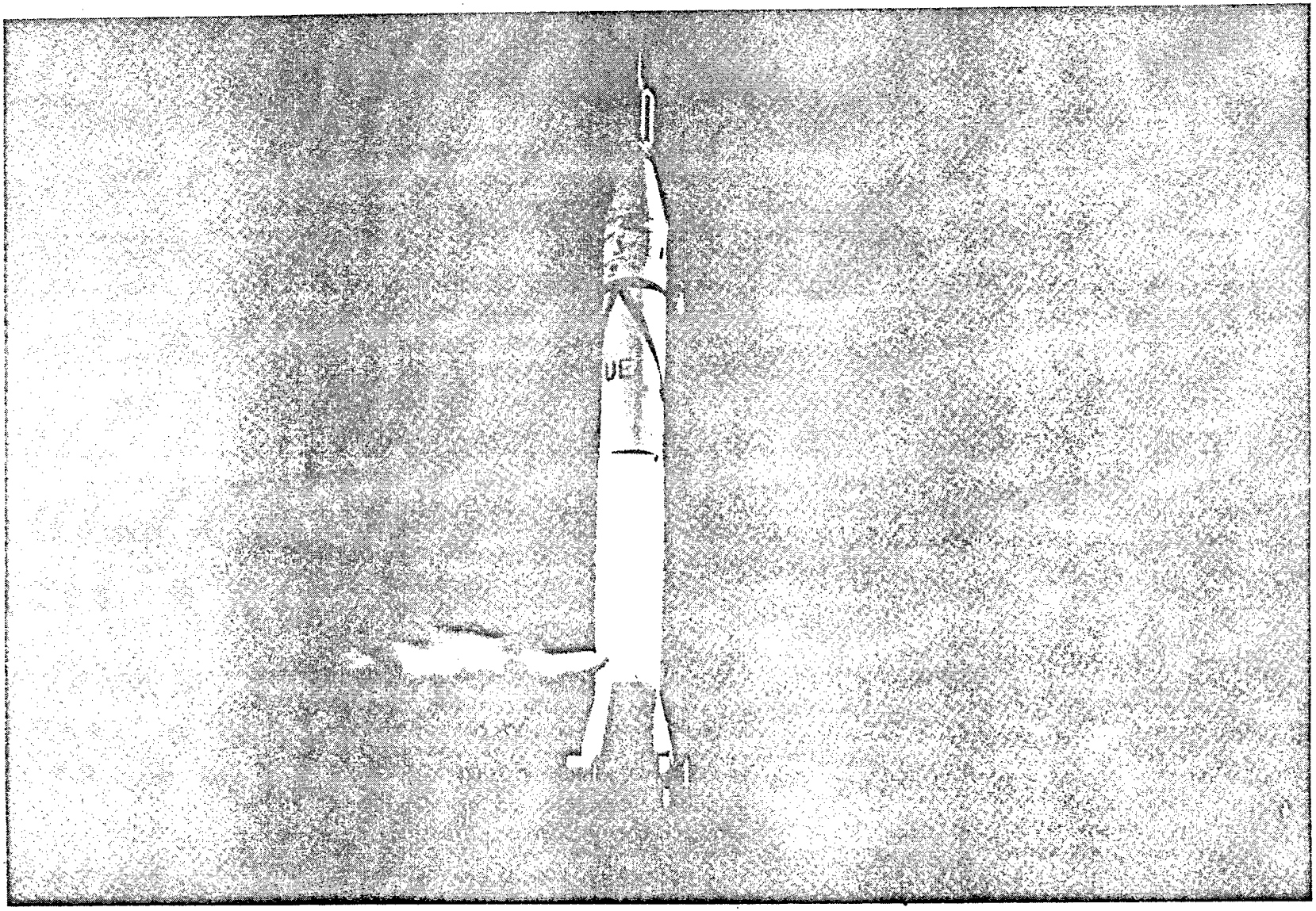
Slide 22



Redstone

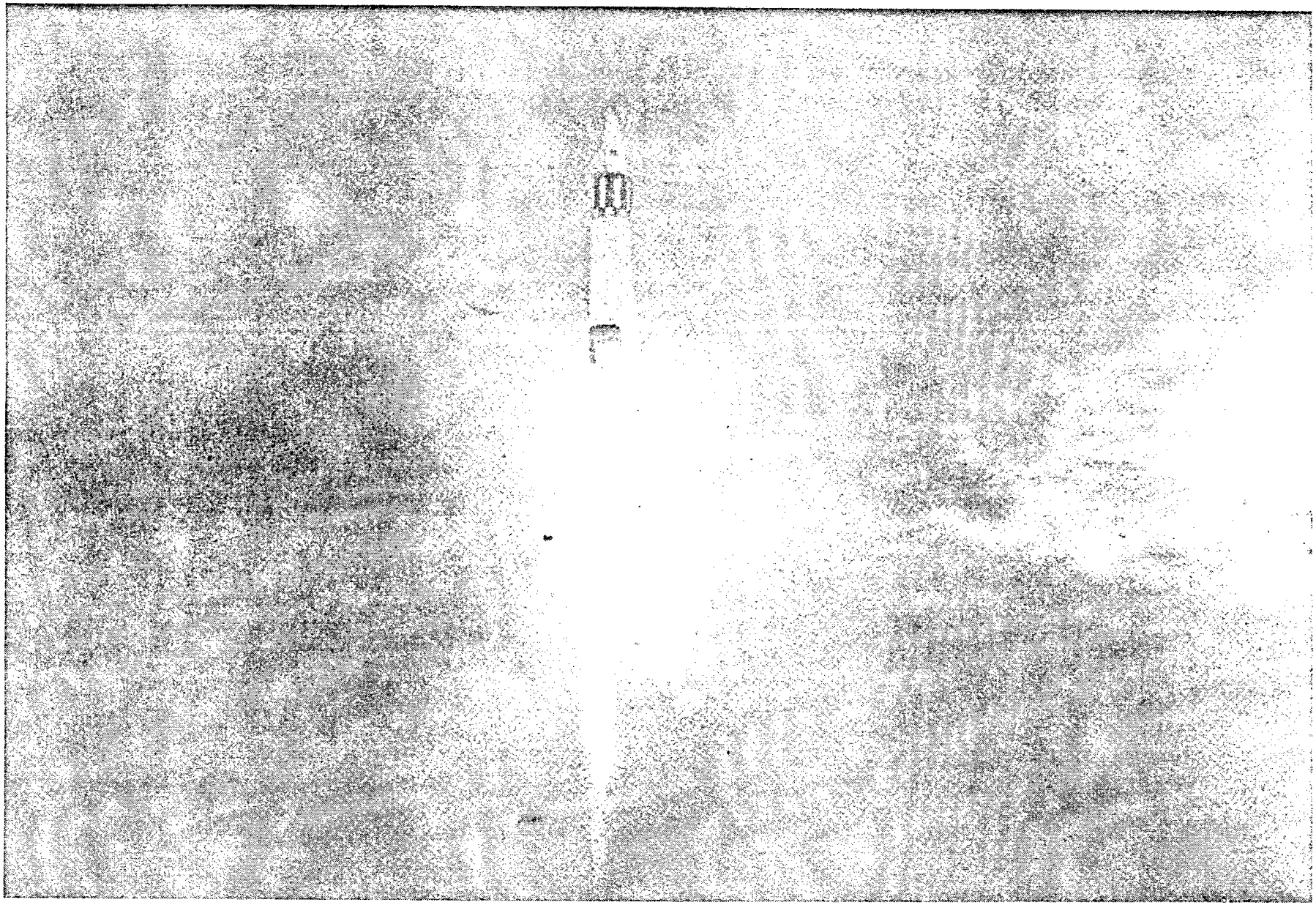
Slide 23

Jupiter C Explorer I



Slide 24

Jupiter



Slide 24

Next is the Army's IRBM Jupiter, and that covers the rocket family.

I want to say something now about contributions in the field of communication. Certainly in the field of transistors, printed circuits and miniature and micro-miniature components, we have entered into a new age in electronic packaging. Truly, this type of component will provide entirely new types of electronic instrumentation. The last I heard was that miniaturization had got down to a point where there are in the order of 600 thousand to a million parts per cubic foot component density, and I also understand that solid state materials are being used actually to build circuits - that is, inductors, resistors, and capacitors - right into the materials; so much can be expected in this area in the future.

Slide 25

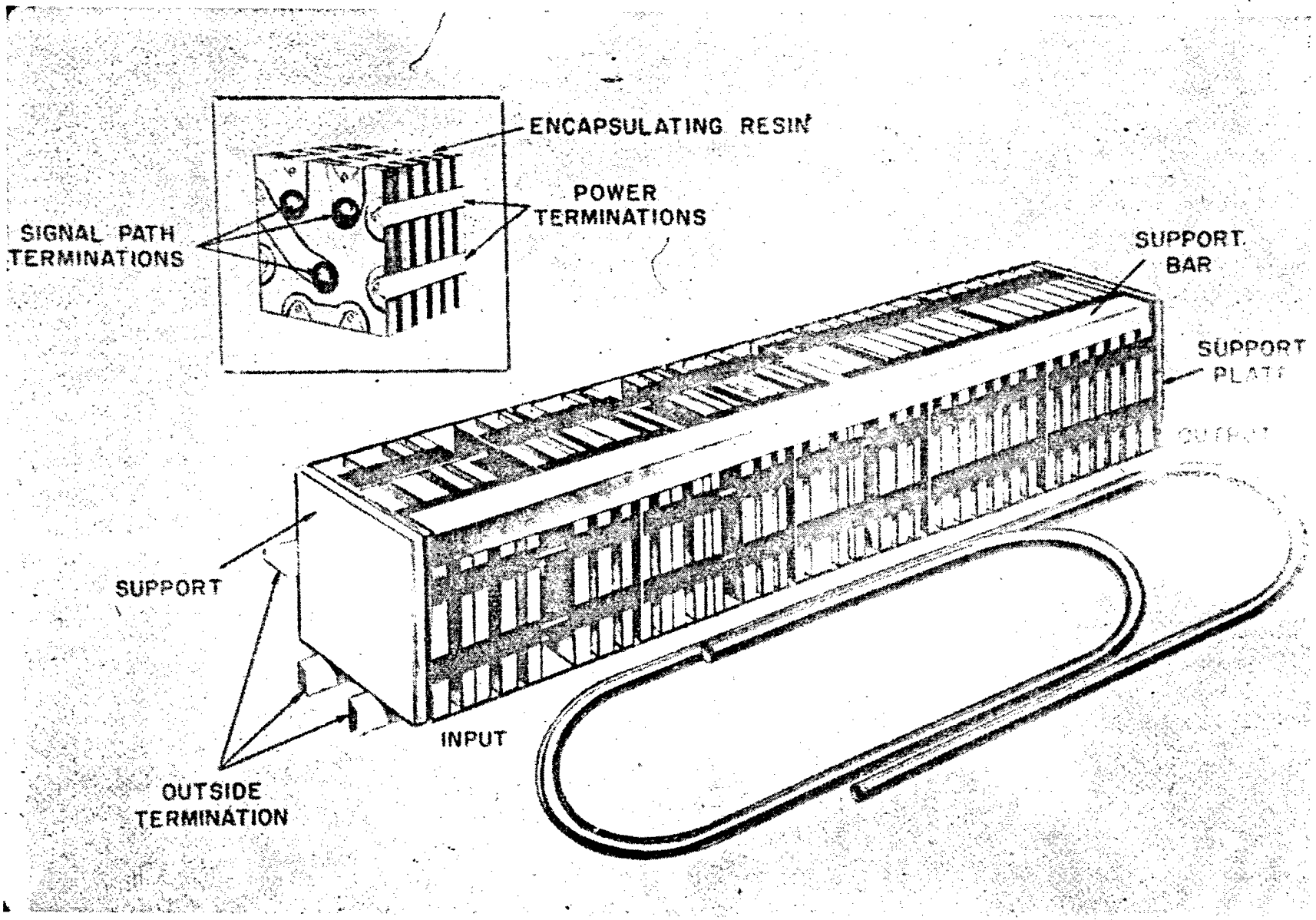
I have several slides which show some of these. The first slide is a miniaturized component as you see it along side of a paper clip. This isn't just one component - it has the entire circuit built inside of it and performs functions such as switching, oscillation and amplification.

Slide 26

The next slide is a picture of a micro-module shown alongside a lump of sugar. There is a stack of circuit elements designed for various functions in the module.

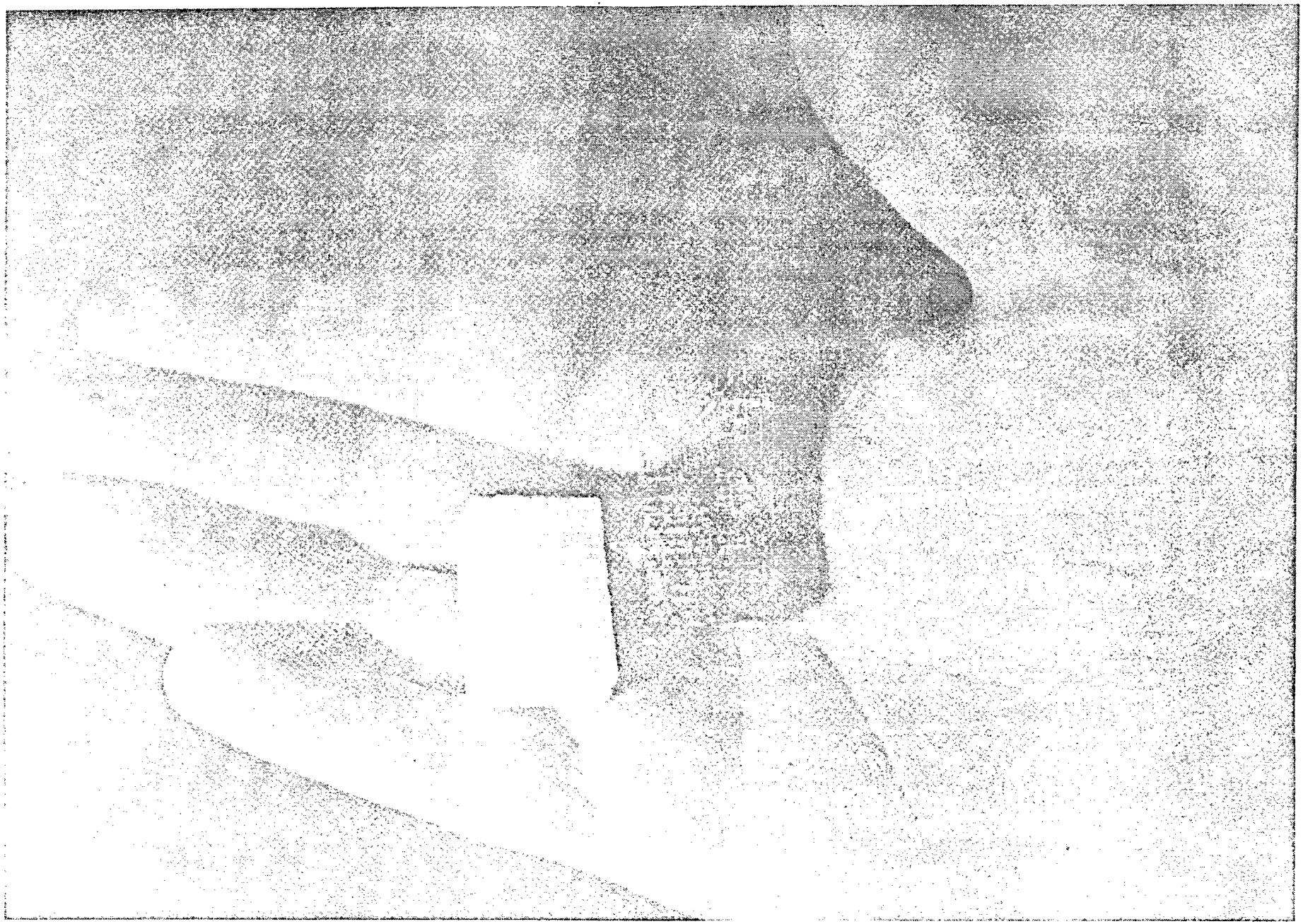
Slide 25

Micro modules



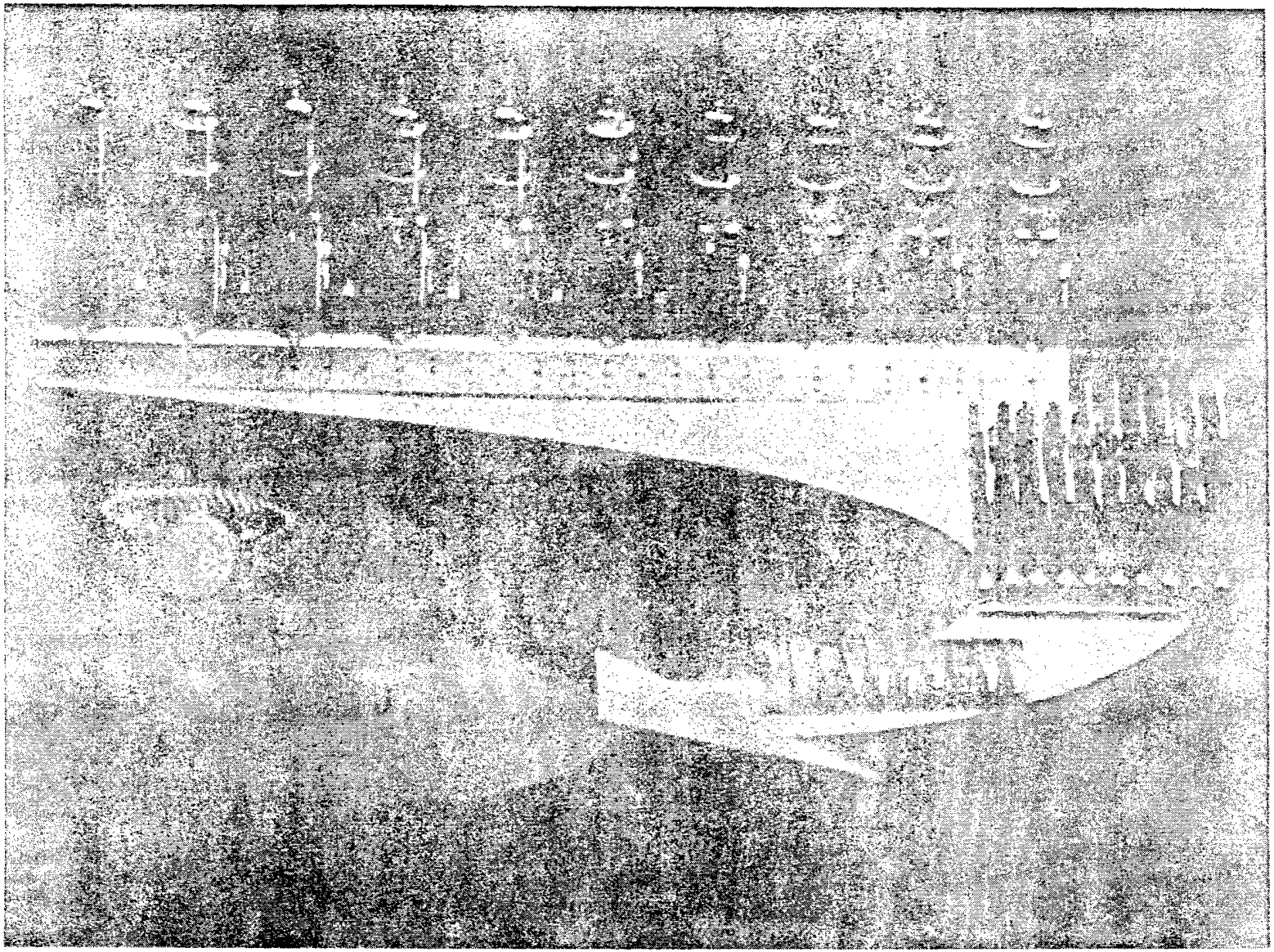
Slide 26

Micro module



Slide 27

Binary Computer

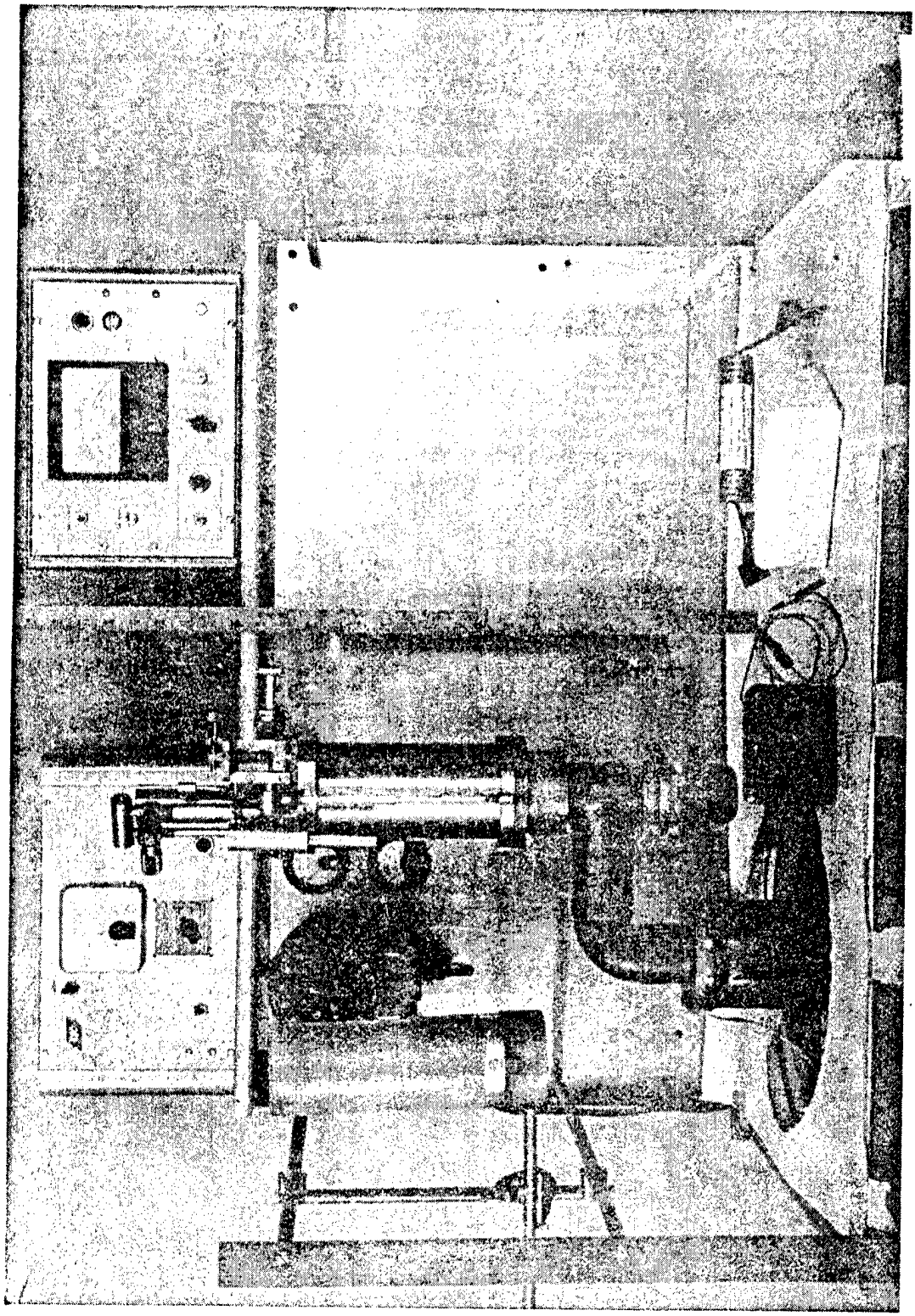


Slide 27

The following slide shows the progress over a number of years - essentially a decade - in a computer. This little "J" down on the front patch represents the item that now does what all of the other ones preceding it did in the past.

I have given you a brief view of the contributions in communications and electronics. I think that one can say equally well that major advances have occurred in radar, in television, and certainly in many fields of science. Much has come from the work being done by the Defense Department, of which the Army is a member.

Certainly one very important field is that of the MASER.



Slide 28
Gaseous Maser

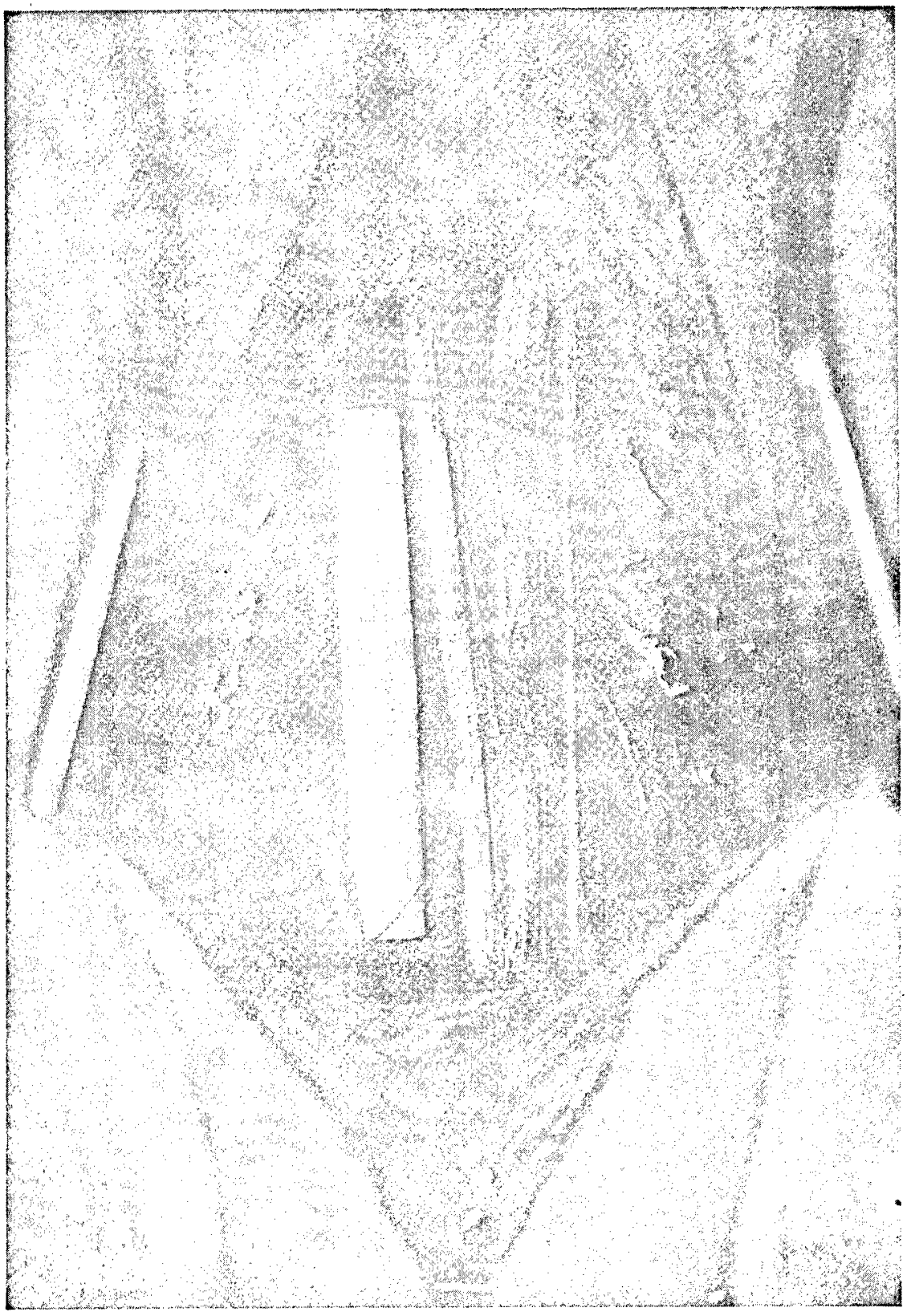
Slide 28

This is a gaseous maser. It came about during the process of studying the structure of hydrogen. For those who might not know, "maser" means "microwave amplification and stimulation of emission radiation." The MASER is an extremely low-noise device and has now made possible developments in radar and communications which are rather phenomenal.

The signal to noise ratio is so high that the maser promises to reduce power requirement from 10 megowatt to 30 kilowatts in a radar application, or reduce antenna size from a 250 ft dish to a 60 ft dish; or increase the range from 3,000 miles to 12,000 miles; or reduce the target cross-section from 150 sq in to about a half sq in; or reduce false alarm rates from one per day to one per nine months. These are all tied in with the fact that the noise in these oscillators or amplifiers is extremely low. Used as a frequency device, the gaseous maser can yeild precisions in the order of one part in 10^9, or reflected in something more popular, would maintain time constant to within one second in 300 years.

And now a few words about the Medical Corps. As you know, they have had a long and rather eminent career in the field of medical research, starting with Walter Reed. They are concerned with medical problems wherever our own American soldiers are. Mass immunization methods are being worked on, as well as yellow fever control. You would be interested to know that when the Asian flu hit this country the Army medical research units had already isolated the virus some three years previously and had determined what were the necessary anti-toxins that would be used to check it. The results of this work were applied to checking the virus when the country was exposed to it.

Important work is being done in the field of nerve repair, using monomolecular films of millipore (and I think this probably means many, many pores). They actually have been able to surround a severed spinal nerve or an optic nerve in an animal with this particular material. There are holes it in which enable nutrients to penetrate through and feed the growing nerves, and actually make it possible for the severed nerves to bridge gaps of the order of several millometers and unite with their proper partners. It has resulted in considerable decrease in scar tissue and should have exceptional success in the repair of many severed nerves.



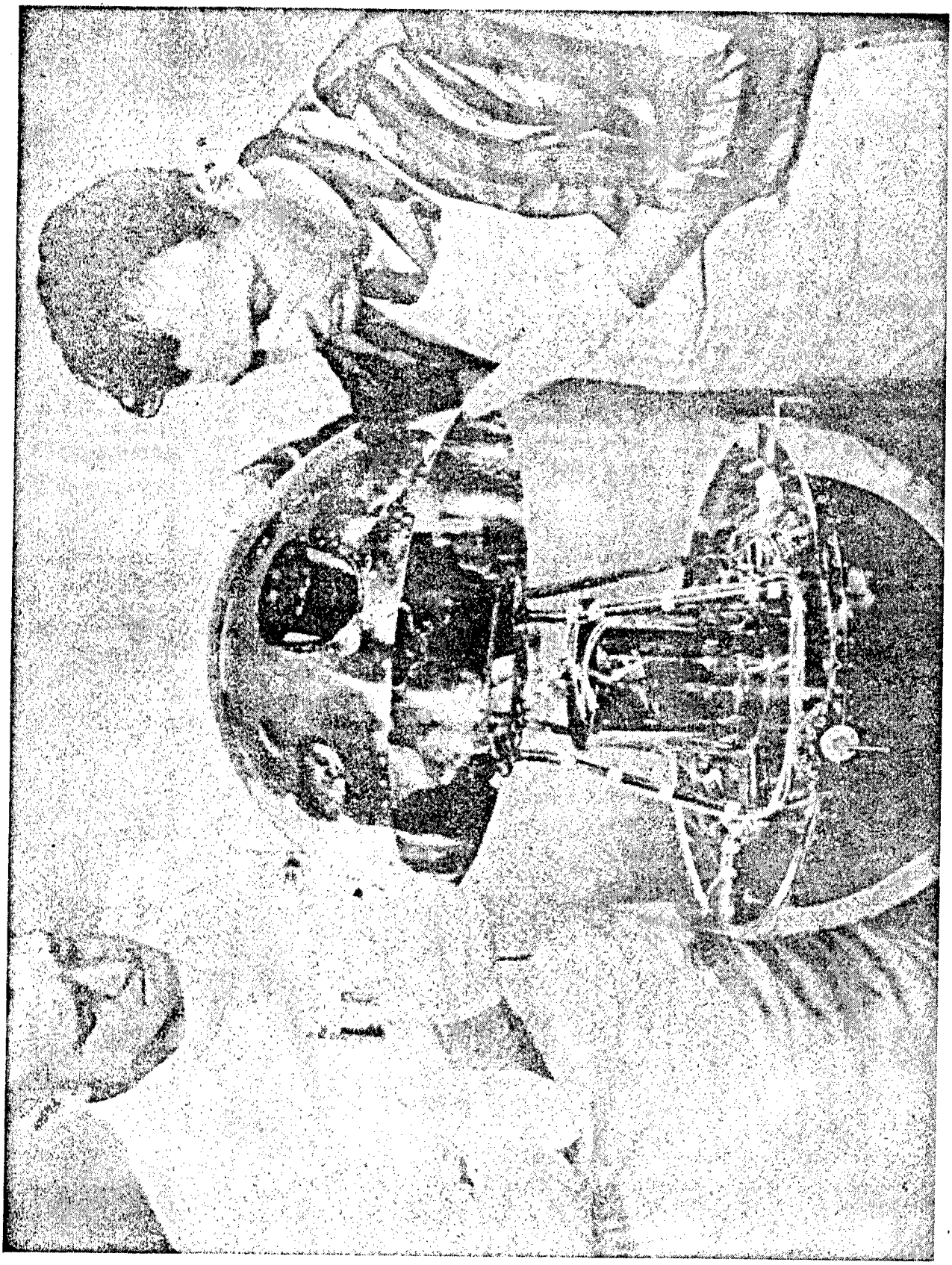
Slide 29
Nerve Repair

Slide 29

This slide shows a picture of that. There you see the severed nerve, and that white material will go around this part and will stay there. The wound will be sewed up and in a short period of time the nerve fibers will reunite.

There is a counterpart to this. This has to do with a new type of bone glue which has been developed by the Walter Reed Army Institute of Research in conjunction with the Hahnaman Medical College and Hospital. This is a polyurethane foam. Actually what happens is when a bone is broken, the break is essentially set, a two-bladed circular saw separated by a proper distance is used to saw out two channels and a piece of bone is taken out of the injured bone, the polyurethane foam is packed into the space, and then the bone is put back in place. In several minutes the glue is hard enough so that it can be chiseled away with a hammer and after sewing up the wound it is possible, within a period of 48 hours, for the patient to walk away. This has been done. So, you see if we combine the gigantic stapling machines which the Russians have developed with this, we have a real good do-it-yourself technique of repair.

In several general areas you know the Army's interest. In the field of infra-red, control instrumentation and photography, much work has been done. Weather prediction comes in for a rather major share of Army research.



Slide 30

Weather satellite

Slide 30

The next slide shows a weather satellite.

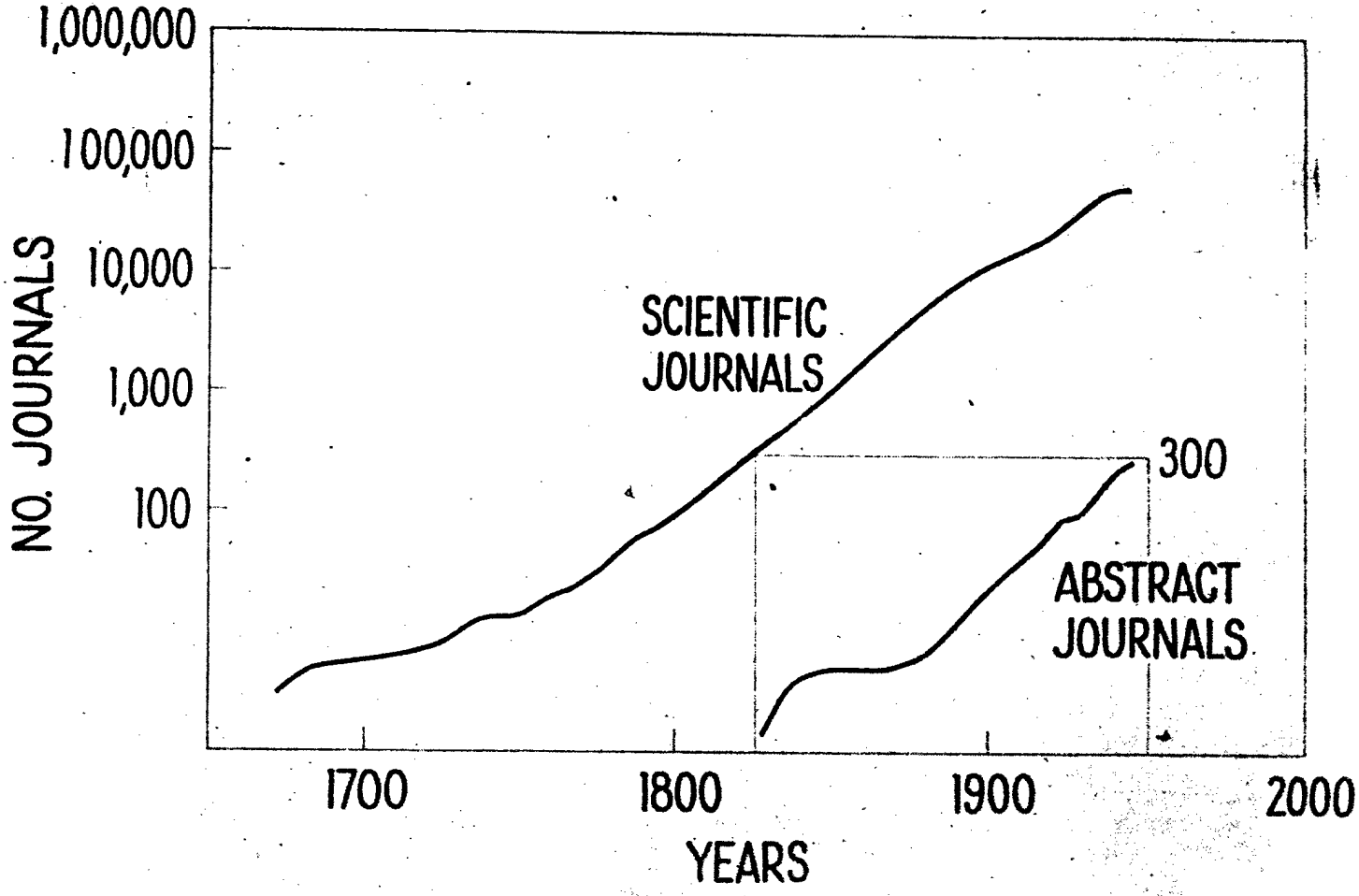
Just a brief summary; I think that one might gather, and I don't have to tell you people, because not only are you representatives of the Army but of industry, and you know that much of the research dollar that goes into defense ultimately finds useful outlets into civilian economy. The taxpayer certainly gets his dollar's worth, we think.

As to the future position of government in research development, it appears as though it will be in it for a long time; first, because it is necessary; secondly, it is part of the way we do things; and thirdly, the growth of knowledge is going along at such a terrific rate that it doesn't appear as though small units in our economy can support the demands that are placed on them. I have two examples here:

Slide 31

Growth of Knowledge

GROWTH OF SCIENTIFIC JOURNALS SINCE 1700

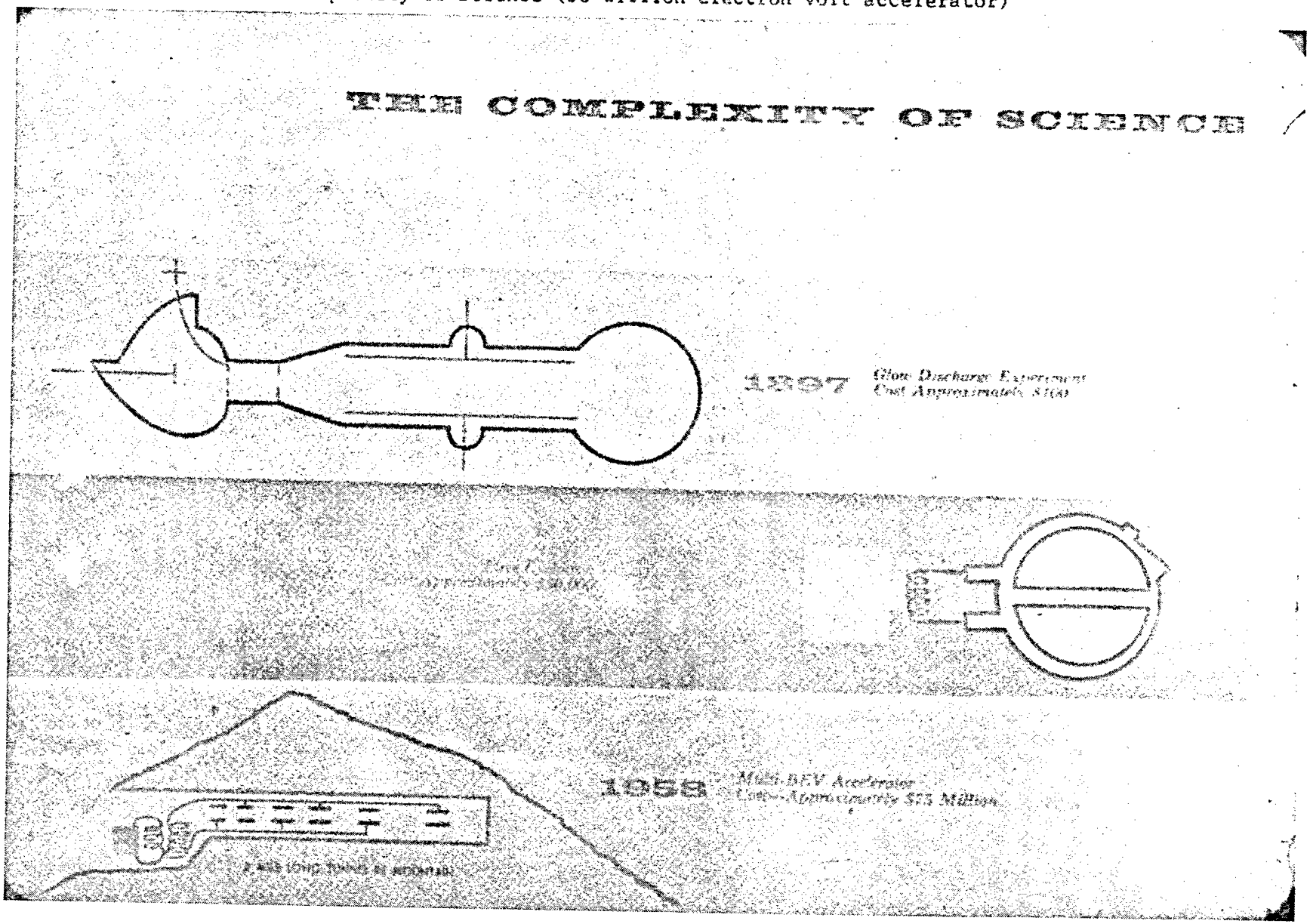


Slide 31

This first slide is a picture of the growth of knowledge, plotted on, as you observe, a semi-log scale from 1700 to the year 2000, and if you look at the scientific journals, you find that the slope is about equivalent to a doubling of knowledge every twelve years. This is on the assumption that there is a one-to-one equivalence between new knowledge and the new data published in the scientific journals. In the field of physics, I understand, it doubles every six years. In the abstract journal field you can see that here the slope of the curve is the same, so that our knowledge is growing so rapidly that even the abstract journals that just report on what is in the scientific journals are experiencing similar problems. Someone said that we renovate our society every two and a half decades, and when one thinks that about 90 percent of the brains that ever existed on the face of the earth are here today, you can well understand it.

Slide 32

Complexity of Science (50 billion electron volt accelerator)



Slide 32

This slide shows the 50 billion electron volt accelerator which is being planned at Stanford Research Institute and is being supported by funds which I understand the President approved; I would like to give you just an indication here of actually what has happened in this complexity of science. In 1897, I think the first experiments by Crooks in a glow-discharge tube cost about $100. In 1934 the first cyclotron, having greater resolution, cost approximately $50,000 and in 1958, the Stanford multi-billion electron volt linear accelerator cost approximately 100 million dollars and required a tunnel two miles long through one of the mountains in California to house it. So you can see tremendous increase in the cost of research. Rather interestingly, the cost per electron volt remained essentially constant through this period. It is 1/10th of a cent per electron volt when you figure it out.

I would like to terminate my discussion by saying that the Army is preparing to meet this rather large challenge of increasing technology and will expect, in the process of doing it, to add even more contributions to the general welfare of the nation.

PREDICTION OF THE RELIABILITY OF COMPLEX SYSTEMS

N. E. Golovin
Advanced Research Projects Agency

The purpose of the following remarks is to outline a point of view toward the reliability of complex systems which we have been developing in ARPA. In so doing, we shall attempt to describe why the problem of predicting the reliability of complex systems is such a difficult one, and hazard some suggestions as to lines of effort which perhaps have not been adequately emphasized because of extensive and somewhat fruitless searches for simple solutions.

First, it is probably advisable to start by defining a few principal terms, some of which have already been used.

By part we will mean the simplest constituent of a group of objects in an assembly of interest. Generally, it is an object which is not normally considered disassemblable into simpler elements. An electronic tube, a transistor, or a capacitor are examples. By a component will be meant an integrated group of parts performing, generally, a simple function in a grouping of similar objects. An instrument, such as a voltmeter or a complete radio receiver or transmitter, can be considered as a component. By a subsystem will be meant an aggregation of components performing a major function in a system. For example, if the system in question is a group of satellites to be used for navigational purposes, a subsystem would be the group of shipborne receivers, computers, and other similar components which transform satellite signals into a ship's latitude and longitude.

The term reliability has been defined in various ways. The following definition is essentially that first introduced by Carhart [1] and seems to have fairly wide acceptance. The term reliability of a system, subsystem, component, or part will be taken to mean the probability that it will perform its required functions, under defined conditions, for a specified operating time. This definition requires that the measure of reliability is to be a number. It presupposes, therefore, that the required functions of the object whose reliability we seek to establish are quantitatively relatable in some way to the numerical measure of its probability for performing them. It also presupposes that means exist for connecting, again quantitatively, the performance of the object to the environmental conditions under which it will operate. As we will see, an important aspect of the difficulty in establishing reliability lies in establishing such quantitative relationships.

Now we are interested here principally in the immediate problem of predicting rather than in evaluating reliability. The former is concerned with assigning a performance probability to a system before it is built, while the latter can be carried out only when at least a system prototype is available for testing. Prediction, therefore, requires estimating the performance probability of a system under conditions when not even a complete design may be available. The importance of prediction is associated

generally with managerial judgments as to a proposed system's practicality or operational usefulness. In major programs, such as the NIKE-ZEUS Missile Defence System, Project Mercury, or a Communications Satellite System, prediction of expected operational reliability must be an integral part of the initial design feasibility study, and, therefore, an essential part of the decision to build or not to build a system prototype for further study. For example, if a communications satellite were to have a predicted mean life (a term which will be defined later) of two months instead of twelve, and its price in orbit runs into the tens of millions, then the associated estimates of the costs of establishing and maintaining a system of say four satellites in effective condition, may well be so great as to cast some doubt on the merits of even a large scale research and development effort. The large costs of such space systems further underline the importance of reasonably accurate reliability prediction because even relatively small differences in expected reliability will correspond to large absolute cost differences. Moreover, systematic reliability analysis in the initial stages of design produces additional engineering inputs for consideration of alternative approaches to an over-all system design. It will be particularly useful for guiding choice of acceptable trade-offs since generally performance, weight, space, cost, and operational reliability have all to be jointly manipulated in attaining an optimized design for the system.

Let's then address ourselves to the situation in which we have a detailed system design before us and see how far we can get in developing a general technique for predicting its reliability. My procedure will be to develop a theoretical approach to the problem interspersed with some comments and comparisons related to current methods in handling arbitrarily complex systems.

From some points of view, the crux of the problem in such an analysis is two-fold: First, the matter of how one defines "failure," and second, how one attempts to construct an expression for the over-all reliability of a system.

Conventionally one considers two types of failure, the so-called "catastrophic" and "degradation" kinds. The first is associated with the sudden, total failure of an object of interest, breakage of the heater element in an electronic tube being an example. The "degradation" type corresponds to gradual deterioration of one or more of an object's characteristic parameters to the point in time where an essential function can no longer be fully performed; for example, the gradually decreasing rate of cathode emission or, more generally, the drift in time of any electronic tube characteristic. In a general analysis, it is difficult to maintain a continuing distinction between these two types of failure, nor is it really necessary. In the subsequent remarks, we will combine these two physically distinct types of failure into one; we will say that an object fails at the time that any of its relevant physical characteristics attain values outside a specified range. Our analysis will try to show how this range must be determined for the general method to be consistent and useful.

As to the manner of constructing an expression for the over-all reliability of a system, the usual procedure is to begin with failure studies of parts and to construct from such data, successively, estimates for the reliability of components, subsystems, and finally of the system as a whole. We will reverse this usual procedure and start with a definition of failure for the system, and then work back through subsystems and components to the data on parts failures. The basic reason for this reversed approach is a somewhat theoretical one; namely, the fact that a part cannot logically be said to have failed unless the over-all system has done so. This means that the definition of part failure must be completely implied by the quantitative definition of what constitutes system failure. This point of view, it should be mentioned, is adopted in MIL-STD-441 for Reliability of Electronic Equipment [2], which suggests that the required performance of system details should be obtained by working back from over-all system functional requirements.

Now the over-all system design must specify how its outputs must fall within certain specified ranges of values if the system's objectives are to be met. The failure of a system to meet design objectives can thus be always unambiguously and quantitatively defined. For example, the transmitter power level in a communications system must be above a definite minimum value, if a specified receiver, at a given location, is to insure a specified, minimum usefulness of delivered information. Furthermore, considering the assembly of distinct subsystems which interact to insure the output characteristics of the system, we can also take as given a set of mathematical relationships which allow calculation of over-all system outputs from the characteristic outputs of the constituent subsystems. This is not an unreasonable assumption. For a design to be at all realizable, such mathematical relationships are either deducible from applicable physical theory or have been empirically established from related prior experience with similar equipments. This is necessarily the case if subsystem, and lower order, nominal performance specifications are to have a rational scientific foundation. As a matter of fact, if such is not clearly the case, it can probably be cogently argued that the state of the applicable theoretical and practical arts does not justify a major system development program.

A key initial point from the reliability analysis point of view is the existence of such a mathematical representation, theoretical or empirical, as a foundation for rational, nominal design specifications. This is the case because such a mathematical representation can be used for constructing, on a computer of adequate capacity, a system simulation program in terms of the output characteristics of all of the system's subsystems. Computer-based system simulation will then allow the systematic study of the effects on over-all system outputs of arbitrary variations in the structure of subsystem output characteristics. The results of this type of investigation, ideally, will be the unambiguous specification of quantitative ranges for subsystem outputs, individually and/or in interdependent groups, which must be maintained if over-all system outputs are to be within the ranges defined by the tolerance requirements for system nonfailure. In this type of Monte Carlo simulation experiments, efficient conduct of the studies would no doubt be aided by experience with statistical experimental design techniques in other fields.

To emphasize the point, the importance of proceeding from the tolerance limits on over-all system outputs to the mathematically implied maximum ranges of allowed variation in subsystem outputs is, principally, that one thereby obtains a valid quantitative definition of subsystem failures. Furthermore, these definitions then permit equally valid specifications of the probabilities of failure of particular subsystems, or of combinations of these into groups, if some are found not to be individually independent with respect to failure. Thus, the probability of failure of a particular independent subsystem is the likelihood that one or more of its outputs fall outside the tolerance limits established as acceptable by such a computer-based simulation. Similarly, the probability of failure of statistically interdependent groups of subsystems is the likelihood that the structure of the groups' outputs to other individual subsystems or groups falls outside the ranges specified by the simulation study. Aside from the quantitative definition of what constitutes subsystem failure, such investigation will thus also have as an inescapable by-product the quantitatively justified grouping of subsystems into statistically independent entities whose probabilities of success or failure can then be validly multiplied together to obtain a measure for the probability of success or failure of the over-all system.

The remainder of the argument should now be clear. In similar fashion, one next treats each subsystem as a mathematically structured assembly of component outputs, and then each component as a mathematically related group of parts outputs. Employing computer simulation, there are then developed quantitative criteria for failures of components and parts, as well as their valid groupings for purposes of combining probabilities of success or failure.

The essential product of these successive simulation studies is then two-fold: (1) We have estimates of the permitted range of parameter values for each part in the system as required for over-all system output acceptability; and (2) we have a quantitatively justified rather than an arbitrary basis for combining part failure probabilities to obtain, successively, such probabilities for components, subsystems, and the over-all system.

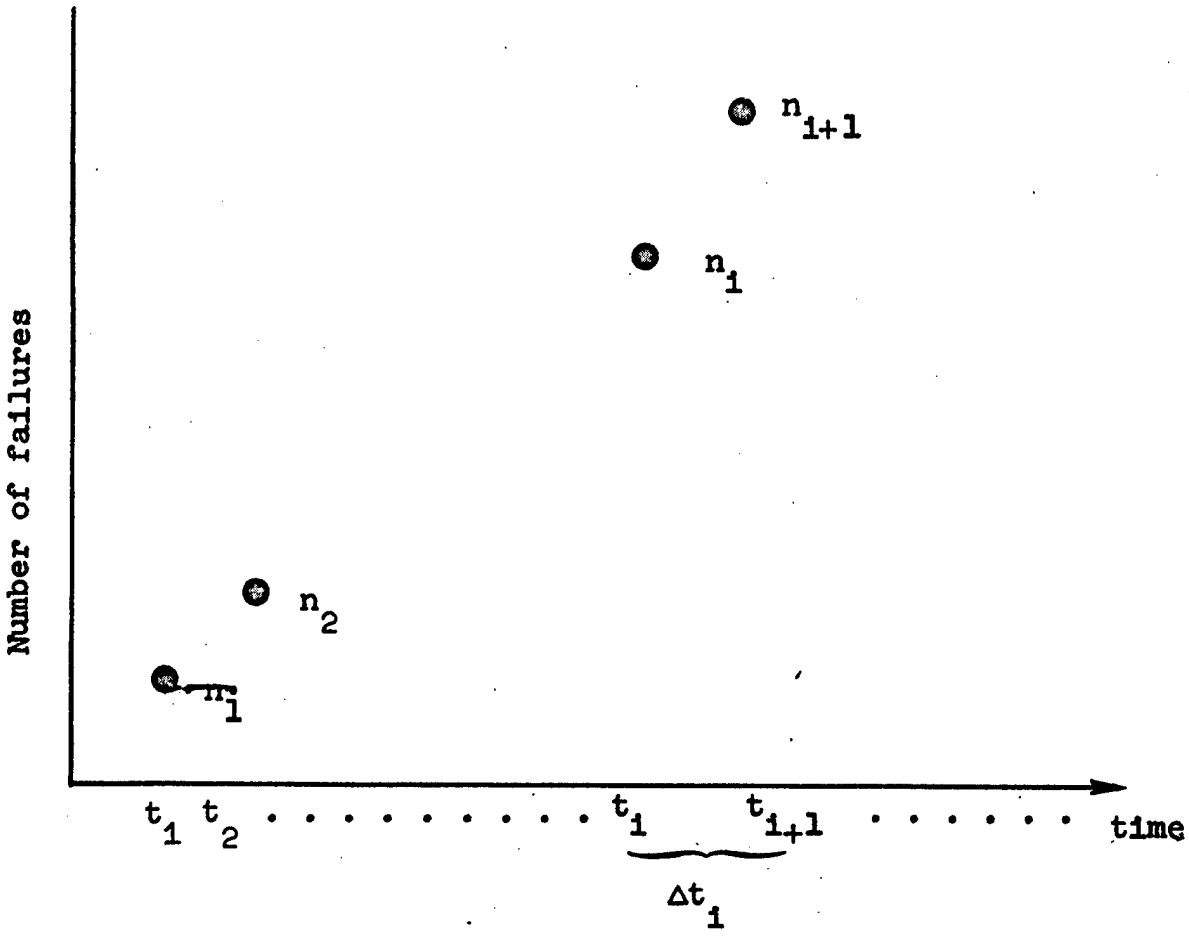
The approach I have outlined has been developing fruitfully, particularly during the last two or three years, in a number of other organizations concerned with complex systems. The White Sands Missile Range, the Rand Corporation, Chance-Vought Aircraft, Convair-Pomona, and the Autonetics Division of North American Aviation have each, in some measure, adopted this approach. At the White Sands Missile Range, for example, the current effort consists of constructing a probabilistic simulation model for the NIKE-HERCULES System in order to develop the technique fully with the view of applying it subsequently to Army missile systems at the design stage [3]. At Autonetics, the technique has been, in part, employed for the analysis and prediction of the reliability of the guidance system for MINUTE-MAN. While I am not familiar with any system to which the approach has been applied in its entirety, it seems that it has sufficiently solid logical merit to grow in importance, particularly in the case of systems in which the commitment to invest in a proposed design is greatly dependent on a

realistic and objectively founded prediction of its reliability, and, furthermore, where the anticipated investments are so great that thorough, and therefore costly, reliability analyses have unquestionable managerial justification.

Let's complete our analysis by turning next to the question of how one develops a generally valid expression for the reliability of an individual part. For each part of the system our computer-simulation investigations have resulted in a quantitative specification of the ranges of acceptable variation in its parameters. Let's assume that we have as many parts as are required for an adequate sample, that we have a clearly defined environment in which the part will be required to maintain its characteristics, and that adequate facilities exist for carrying out a life test of the sample in such an environment. Incidentally, of all the assumptions so far made in this discussion, these last are among the most unreasonable. Usually, at the design stage of a system, many parts do not yet exist, the environment in which they must operate is not clearly established, and testing facilities allowing study of their performance under a realistic reproduction of the anticipated environment are almost never available.

We can test this sample until all of its members fail, and accumulate our results in the way shown in figure #1.

Figure 1



N = Number of parts initially in sample

$\Delta n_i = n_{i+1} - n_i$ = Number of failures in Δt_i

$N - n_i$ = Number of parts operational at time t_i

$\frac{\Delta n_i}{N - n_i}$ = Probability of failure during Δt_i

$1 - \frac{\Delta n_i}{N - n_i}$ = Probability of survival during Δt_i

With such information available, we can then carry out the calculation shown in the next two figures:

Figure 2

$R(t_i) \equiv$ Probability that a part survives time t_i

$R(t_i+\Delta t_i) \equiv$ Probability that a part survives time $t_i+\Delta t_i$

Then

$$R(t_i+\Delta t_i) = R(t_i)\left[1 - \frac{\Delta n_i}{N-n_i}\right], \text{ or}$$

$$-\Delta R(t_i) = R(t_i)\left(\frac{\Delta n_i}{N-n_i}\right), \text{ or}$$

on dividing through by Δt_i ,

$$\frac{-R(t_i)}{\Delta t_i} \cdot \frac{1}{R(t_i)} = \left(\frac{\Delta n_i}{N-n_i}\right)\frac{1}{\Delta t_i}$$

Introduce the definition of $\lambda'(t_i)$:

$$\left(\frac{\Delta n_i}{N-n_i}\right)\frac{1}{\Delta t_i} \equiv \lambda'(t_i) ,$$

where $\lambda'(t_i)$ is the probability of failure per unit time given by the sample for the interval Δt_i. We can then write:

$$-\frac{\Delta R(t_i)}{R(t_i)} = \lambda'(t_i)\,\Delta t_i$$

Figure 3

Making the usual assumptions we can then pass to a differential relationship of the form:

$$-\frac{dR(t)}{R(t)} = \lambda(t)dt ,$$

which, on integration becomes:

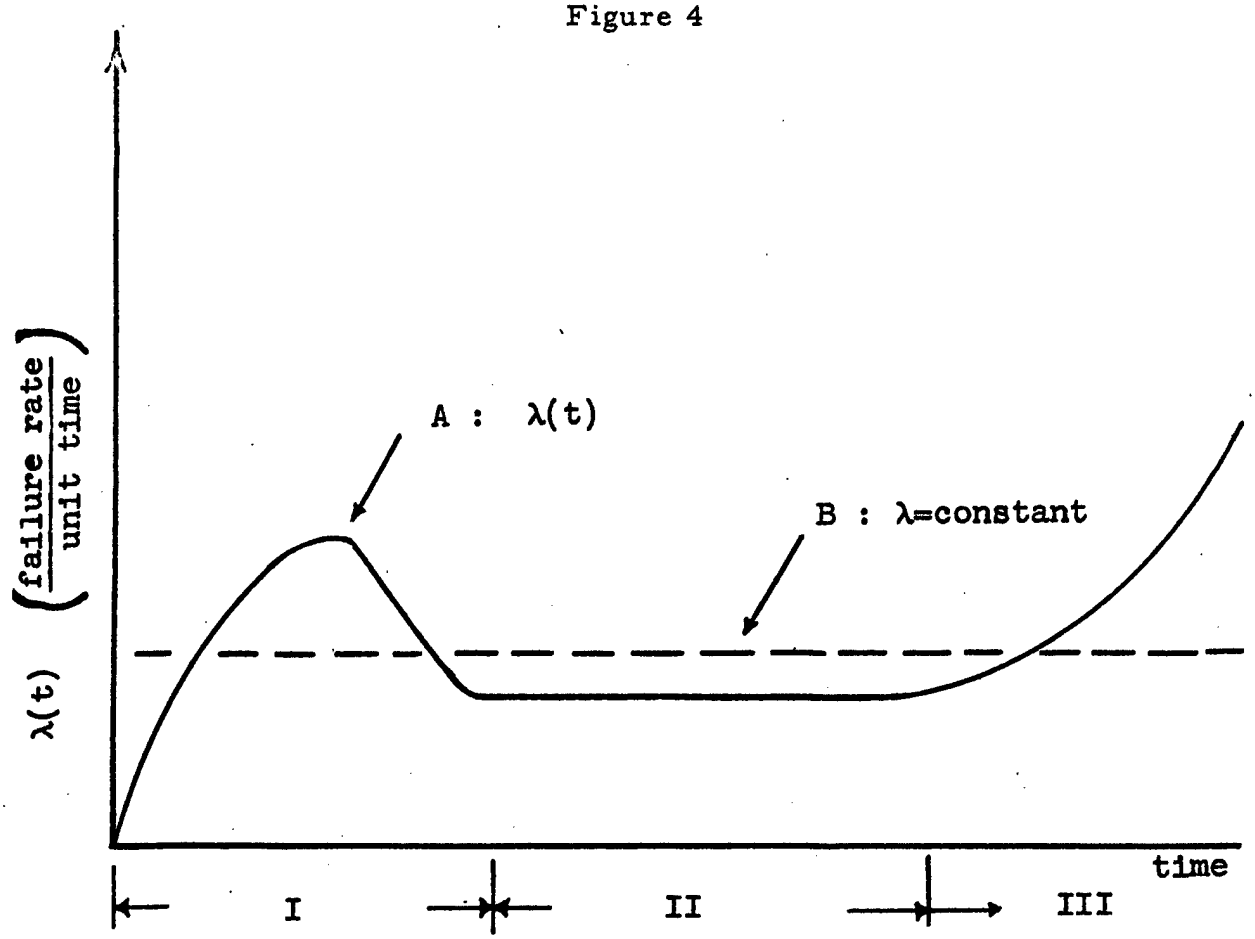
$$R(t) = R(o)e^{-\int_o^t \lambda(t)dt}$$

Here R(o) is the probability that the part is functional at the time the system begins its operation. Practically speaking this can hardly ever be taken as unity, but may be assumed close to this value. So we usually write:

$$R(t) = e^{-\int_o^t \lambda(t)dt}$$

The function $\lambda(t)$, let's call it "the part λ characteristic" may be arbitrary in character. In general, it is supposed to have the form of curve "A" in the following figure:

Figure 4



Region I : "Infant mortality" period

II : Mature operating life period

III : Rapid deterioration period

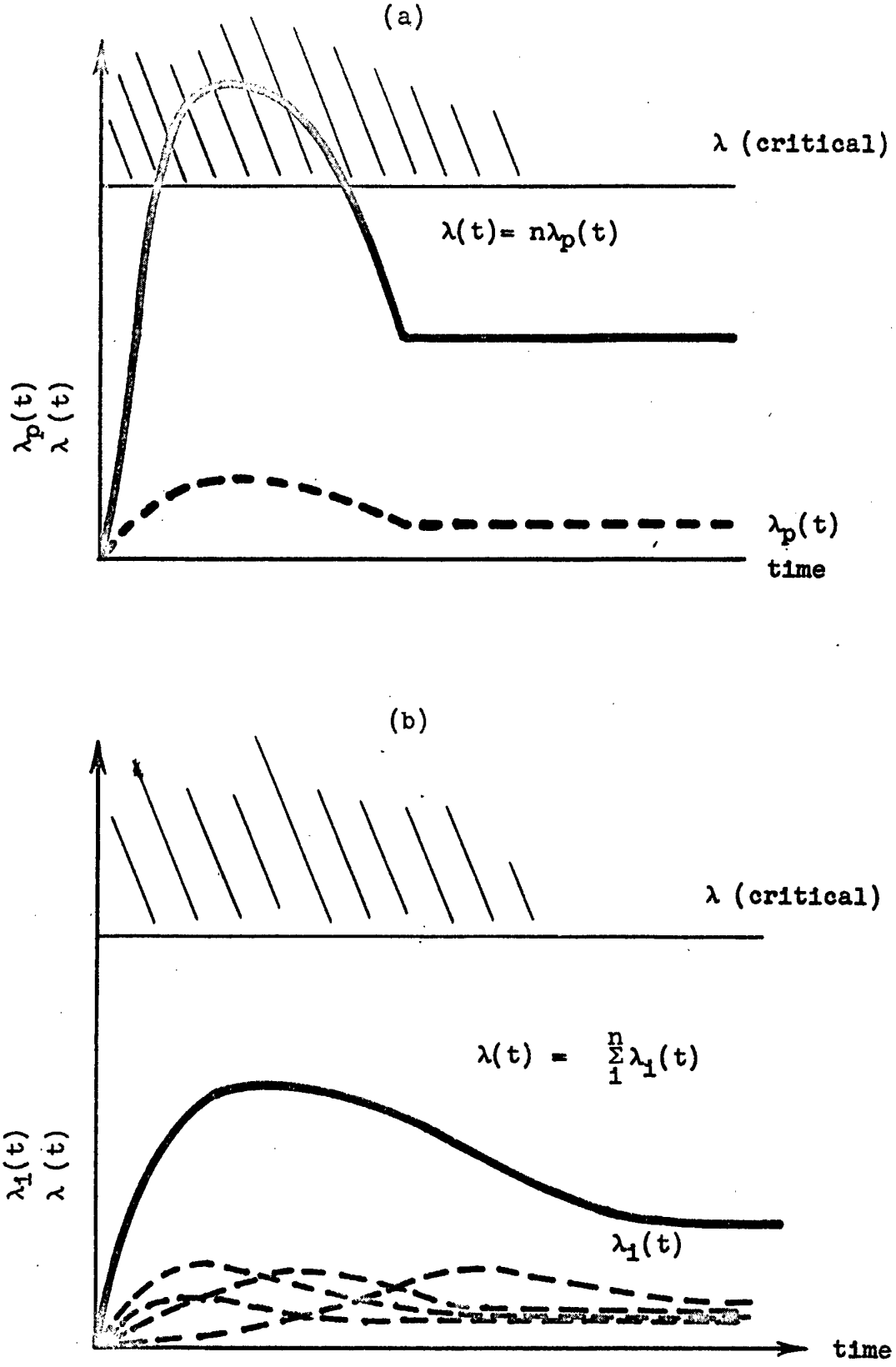
The straight line "B" is an average constant λ characteristic which, largely for the sake of simplicity in applications, is frequently assumed as applicable to most parts and components, for the purpose of taking, so to speak, a "first cut" at describing the corresponding reliability functions.

Substantial effort has gone into finding analytically tractable expressions for the part λ characteristic, or for its reciprocal defined as "the mean time to failure." The usual practices are to assume either that λ is constant, as has been mentioned, or that the part's mean time to failure is normally distributed about some average value with an appropriately chosen variance. There are many applications in which such simplified distributions are useful. However, it must be kept in mind that when many failure rates have to be added together to get a composite rate, the errors in such rates are also added. Accordingly, particularly in the case of complex systems, numerical methods allowing the use of actual rather than assumed part failure characteristics should be employed if at all possible. Additional reasons for care in this connection are suggested by the following:

1. In the first place, if a part is to be employed, without a prior "burn-in" period, in a component in which it is duplicated a large number of times, the superposition of the "infant mortality" periods may result in a total failure rate, for an appreciable time, which is not acceptable for the component or subsystem. This effect is shown in figure #5(a).

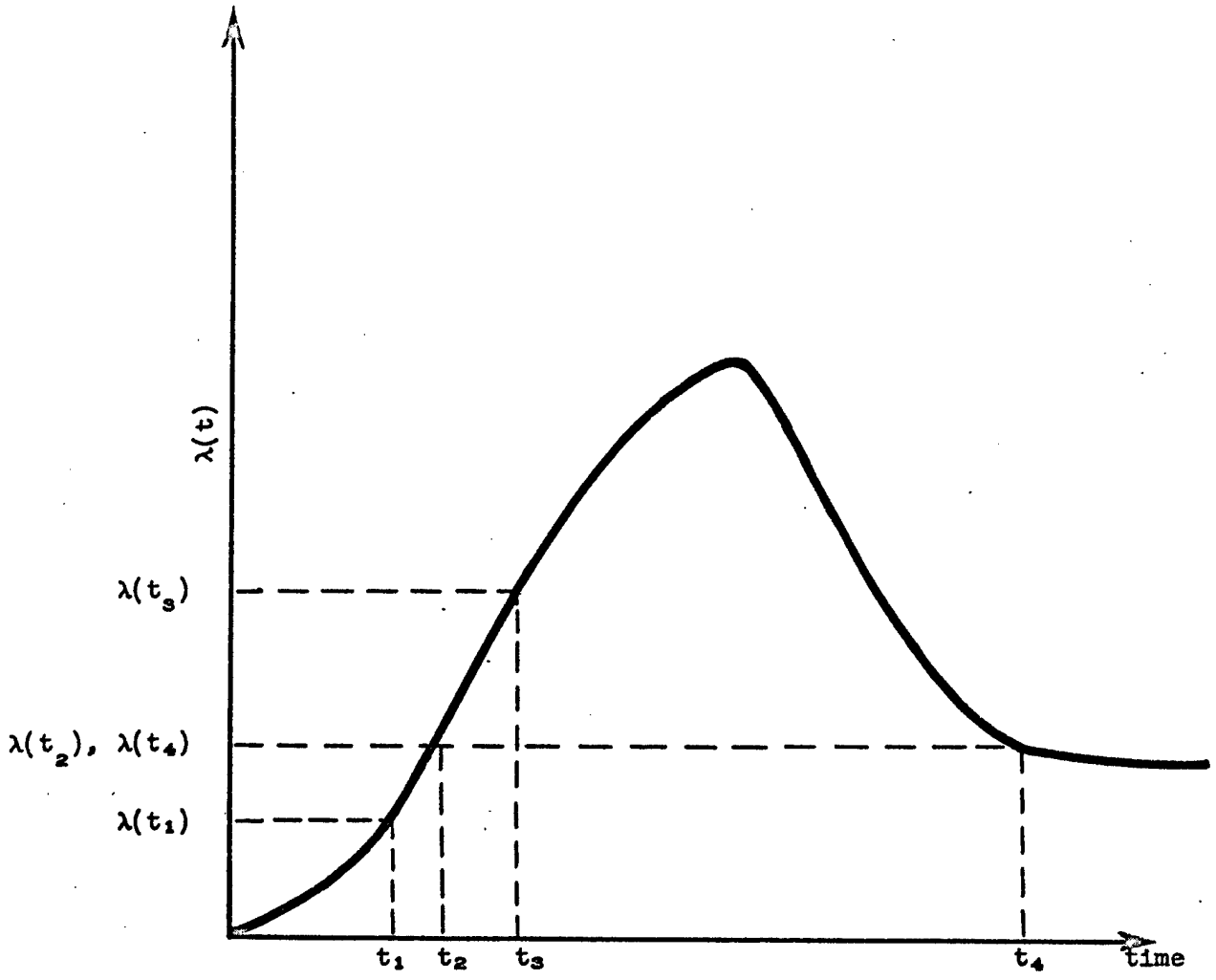
2. Secondly, if a number of different parts, with varying λ characteristics, are appropriately employed in a single component, it may be possible to design easily arrangements in critical circumstances that lead to a component λ characteristic which has desired form or a maximum desired failure rate level. How this can be done is shown in figure #5(b).

Figure 5


(a)
λ (critical)
λ(t)= nλp(t)
λp(t)
λ (t)
λp(t)
time
(b)
λ (critical)
λ(t) = Σ λi(t)
λi(t)
λ (t)
λi(t)
time

3. Thirdly, if a part is to be "burnt-in" prior to use, the required "burn-in" period cannot really be adequately established without constructing the time dependent λ characteristic. This can be seen readily in the next figure (#6). If the system operating period is t_1, for example, a "burn-in" period in unnecessary and even harmful since $\lambda(t_1) < \lambda(t_4)$. If the system operating period is t_2 or greater, a "burn-in" period is desirable.

Figure 6



The last three figures, and some of the associated arguments, have been taken from Druzhinin's article [4] in the book Reliability of Radio-Electronic Apparatus, published in 1958 by "Soviet Radio." This book, incidentally, is the first of promised annual publications of collections of research papers in the field. Apparently in this, as in so many other fields, U.S.S.R. technical organizations have initiated a systematic, broad-based approach. The National Bureau of Standards, in general, and Joan Rosenblatt, in particular, are to be thanked for their initiative in providing translations of some of the more important Russian papers in the reliability area.

Having established λ characteristics for all parts in the system, we can then directly employ the results of the previous analysis for systematic construction of the reliability functions for components, subsystems, and the over-all system.

The procedure can be illustrated by the argument on the next figure (#7), where $R_p(t)$ is the "part" reliability function:

Figure 7

$$R_p(t) = e^{-\int_0^t \lambda_p(t)dt}$$

For a component of n "independent" parts:

$$\lambda_c(t) = \sum_{i=1}^{n} \lambda_i(t) \quad , \text{ and}$$

$$R_c(t) = e^{-\int_0^t \left(\sum_{i=1}^{n} \lambda_i(t)\right)dt}$$

If there are K identical parts in a simple redundant arrangement, the groups reliability function is:

$$R_g(t) = 1 - \left[1-R_p(t)\right]^k \quad ,$$

where $R_p(t)$ has the form shown above.

The component reliability function, $R_c(t)$, [assuming (n-k) "independent" parts and a single group of k parts in a redundant arrangement] is then:

$$R_c(t) = R_g(t)e^{-\int_0^t \left(\sum_{i=1}^{n-k} \lambda_i(t)\right)dt}$$

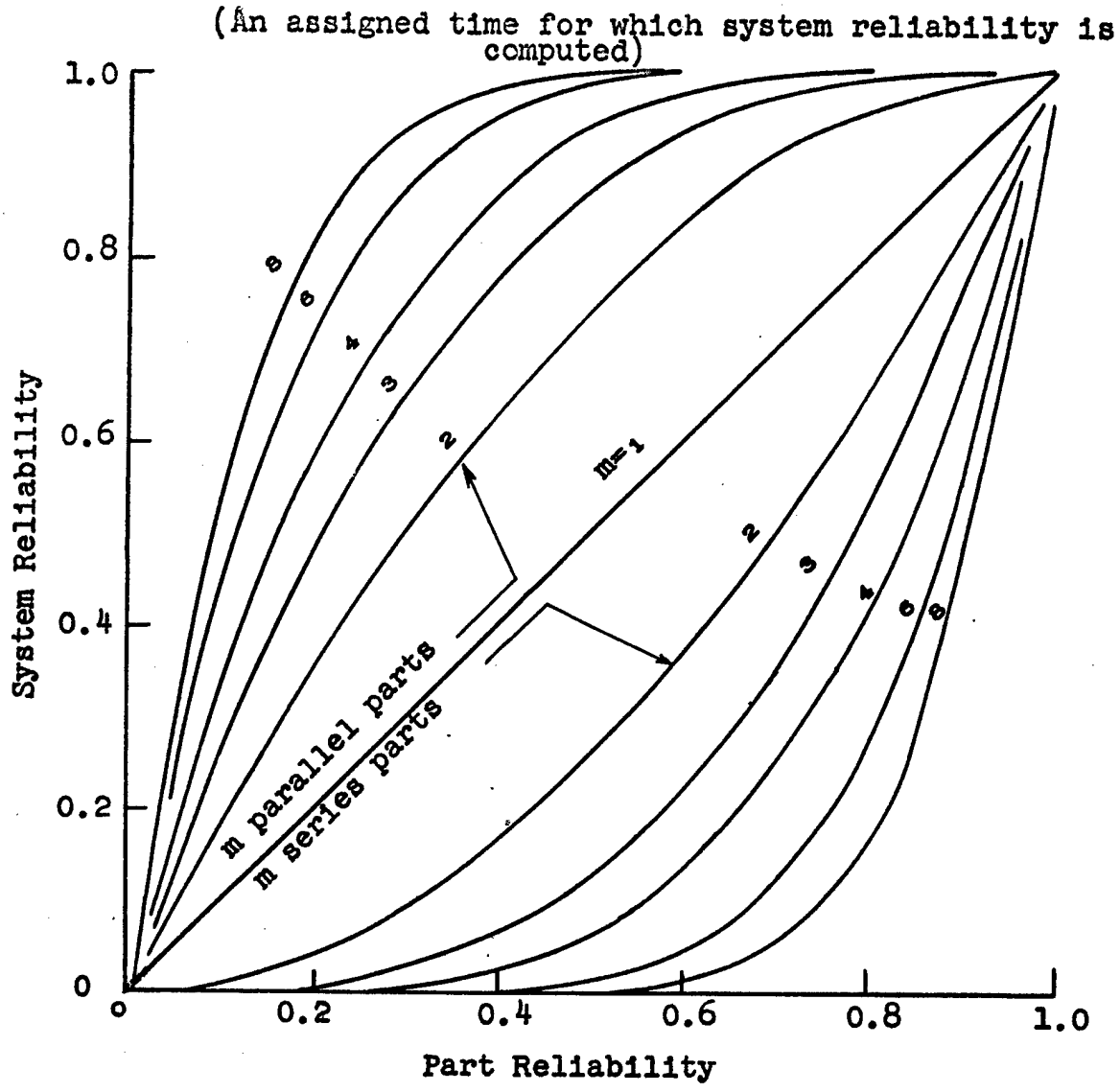
We are now familiar with the reliability function for a single part. If a component, for example, consists of n parts, which our analysis has shown to have independent failure probabilities, then the λ characteristic for it is simply the sum of the part characteristics and its reliability function is as shown on the figure. In the general case, $\lambda_c(t)$ must be obtained by the detailed superposition process of the type previously shown in figure #5.

On the other hand, returning to figure #7, if the component has been found, through our analysis, to have a group of parts which must be treated as an entity with respect to independence of failure probability in relation to the other parts, the reliability function for the group must be built up in accord with the logical relations found for the parts in the group. Probably the simplest case of this sort occurs when the group's parts merely provide functional redundancy. In such a case, the group and component reliability functions can be obtained as is shown on the figure.

The argument for other components, for the subsystems, and the over-all system then proceeds in an analagous way.

Here it should also be mentioned that another important by-product of the general method outlined is that if the resultant over-all system reliability is found to be, for example, unsatisfactorily low, a firm basis has been established already for evaluating the regions of the system where increased part or component reliability, or the employment of redundancy, will be most effective in raising the over-all reliability of the system. Incidentally, the relative values of improving part reliability and redundance, as well as the reliability degradation due to multiplying parts in series, can be inferred from the following figure (#8). In computing these curves, the parts are assumed identical and their reliability is assumed to follow the exponential law.

Figure 8

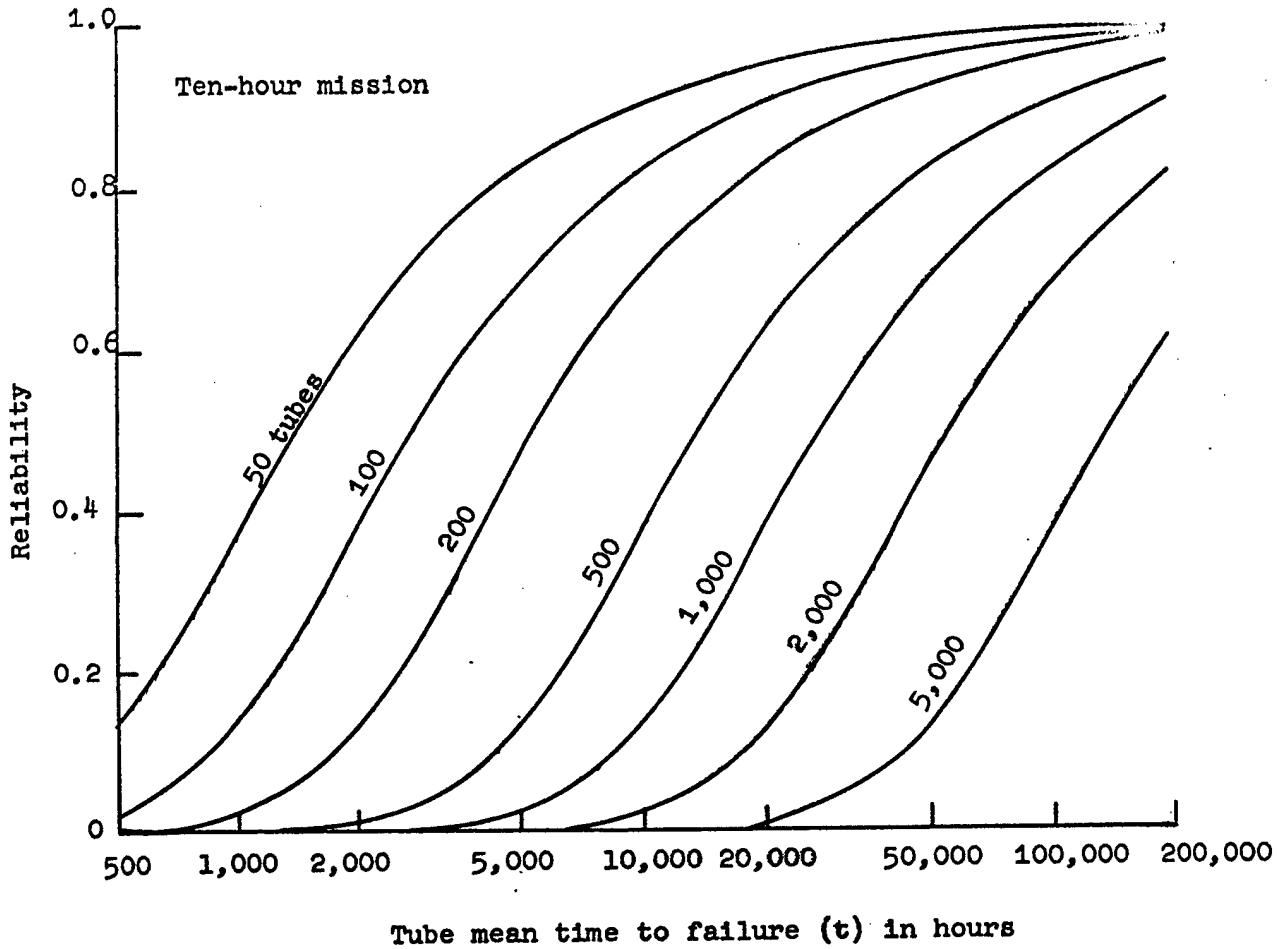


Reliability of a system of m identical parts

The interaction of increasing complexity and part reliability is striking, as shown in the next figure (#9). This is based again on the assumption that all parts have identified constant λ characteristics and are independent in their effect on over-all system reliability.

The last two figures are taken from R. R. Carhart's Rand Corporation Report, "A Survey of the Current Status of the Electronic Reliability Problem."

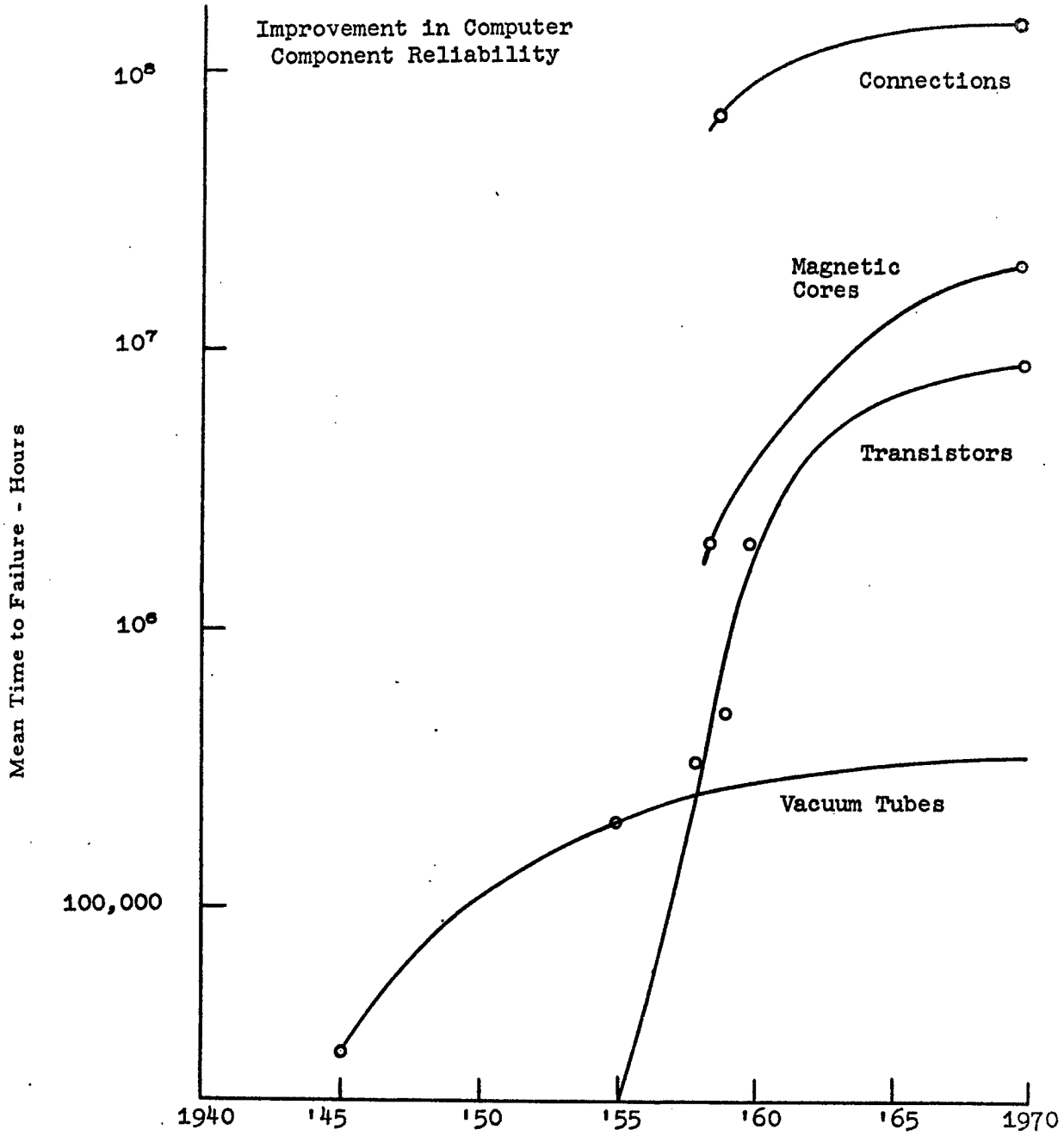
Figure 9



Reliability for 10-hour mission vs tube mean time to failure

In discussing part testing as a part of the job of predicting the reliability of systems in early stages of development, it was implied that the parts to be used in the system may not actually be available for tests to determine their λ characteristics. Of course, if such restrictions exist, there is little choice but to use available failure rate information for similar parts previously tested or used under closely related environmental conditions. However, this is an extremely dangerous procedure, at least for systems whose development cycles extend over several years. This is the case because technology is advancing so rapidly in some fields that errors of several orders of magnitude are possible unless careful and explicit allowance is made for the changing state-of-the-art. The following figure (#10), taken from a recent IBM report [5], shows what is expected to happen in a field of major importance to military applications -- that of computers.

Figure 10


Improvement in Computer
Component Reliability
Mean Time to Failure - Hours
10⁸
10⁷
10⁶
100,000
Connections
Magnetic
Cores
Transistors
Vacuum Tubes
1940
'45
'50
'55
'60
'65
1970

It should be pointed out here that the indicated procedure is not offered, of course, as a panacea. It is being merely suggested that its employment will furnish a very useful and powerful tool for integration into conventional design practices. Nor has any mention been made of the pervasive and insidious influence on system reliability of various human factors throughout both the developmental and operational phases of a system's life.

Looking back over what has been said, it is clear that an obvious aspect of the problem of reliability prediction has been omitted; namely, a discussion of who is it that is going to make the analysis and take the responsibility for the ad hoc assumptions and simplifications that usually need to be made in applying any theoretical structure to a practical situation. The question is far from trivial because it is possible to get from highly responsible and competent groups, as has been already mentioned, estimates of reliability which differ by a factor of more than 10. An adequate treatment of this question might well warrant a time comparable to that which we have already spent. However, a number of assertions are rather readily in order, particularly for the type of approach which has been outlined.

1. In the first place, a key requirement in our point of view is clear articulation of the system design into associated mathematical structures and computer-simulation schemes. Such an undertaking must necessarily be undertaken either by the design staff itself or by a group otherwise living with the job. The requirement for this association is not just a matter of the complexity of the task, but also that the design program itself will benefit enormously from a thorough-going application of the suggested procedure.

2. In the second place, adequate resources must be provided for part development, procurement, and testing; sufficiently realistic environmental conditions must be available for the program, and means must exist for taking advantage both of information available on part and component performance from other contexts, and of changing technology which rapidly makes most extant information quickly obsolete. These tasks, most conveniently, must also be closely associated with the engineering group responsible for design.

3. In the third place, strong motives must exist for a realistic approach to the problem of reliability prediction if useful results are to be made available. This is clearly the case because of the frequency of situations, particularly in the case of complex systems, where no useful data is available and judgments must be substituted. Strong motivations to conservative and realistic prediction obviously can be found only if financial losses rather than gains are to be expected from lack of realism. These considerations suggest that the guidance and technical direction necessary for an adequate and realistic reliability prediction program cannot be

> expected usually to be found ready-made in the group responsible for justifying the feasibility and usefulness of a design. This is particularly the case with complex systems for which it is out of the question to require a serial post-review of the designer's estimates -- simply because the task is too big and will probably require more time than can be afforded in postposing decisions to accept or not accept a given design.

There is, of course, no question that the system design contractor must have a competent reliability analysis group and that it should, at least administratively, not be under the direct management control of the design group itself. However, the above remarks suggest, further, particularly in the case of government procurement actions for new complex systems in which matters of operational reliability are of basic importance, that somewhat more attention than may have been customary in the past be given to developing reliability analysis and prediction programs coincident in time with the beginning of a system design. As a matter of fact, it can probably be persuasively argued that a thorough-going, coincident effort is likely to be not only more fruitful but also, in the final summing up, including operational phases, much less costly.

Also, it seems fairly clear that such reliability analysis and prediction programs should proceed either under the direct, in-house guidance of the government or be conducted under such guidance assisted by adequately motivated contractors not themselves committed to major R&D management or hardware programs in the systems being so studied.

REFERENCES

[1] R. R. Carhart, A Survey of the Current Status of the Electronic Reliability Problem, Rand Corporation Report, RM-1131, August 14, 1953.

[2] MIL-STD-441, Reliability of Military Electronic Equipment, 20 June 1958, Office of the Secretary of Defense, Supply and Logistics.

[3] Exploratory Conference on Missile Model Design for Reliability Prediction, Report of the 3rd Meeting, White Sands Missile Range, 23 April 1959.

[4] G. B. Druzhinin, On Methods of Calculating the Reliability of Systems, Soviet Radio, Moscow, 1958.

[5] Digital Computer Characteristics for Space Applications, IBM Corporation, Federal Systems Division, Oswego, New York, June 9, 1959.

ON THE REPEATED-MEASUREMENTS DESIGN IN BIOLOGICAL EXPERIMENTS

Ardie Lubin
Walter Reed Institute of Research

SOME DIFFICULTIES IN USING THE REPEATED MEASUREMENTS DESIGN. The phrase "repeated measurements design" is used to characterize those experiments where each subject* is tested more than once. Usually this is done to increase the precision of the experiment by eliminating the between-subjects deviance from the estimate of error deviance. Often it is done to avoid multiplying the number of subjects used in the experiment.

The main emphasis of this paper will be on the design where each subject receives only one treatment, applied repeatedly over a period of time, and the chief interest is in the chronic effect of the treatment. An example of such a design would be a drug experiment where each subject is given a constant drug dose every day and tested periodically.

The rest of the paper will discuss the multiple treatment cross-over design in which each subject receives a single treatment for a fixed unit of time, but is changed to a different treatment wherever a new time unit starts. A common example would be a drug experiment where a subject might be on drug A the first week, drug B the second week, and so on. The separate effect of each drug is then estimated from the results.

The purpose of this paper is to point out that: a) any repeated measurements on the same organism will in general exhibit statistical dependence; therefore multivariate analysis of variance rather than univariate analysis of variance is appropriate, and b) all standard cross-over designs assume that the carryover effect of a treatment on a succeeding treatment is constant and does not depend on the nature of the succeeding treatment, i.e., carryover is additive and does not interact with succeeding treatments.

Most of this paper is concerned with possible experimental and statistical answers to the questions which arise when dependent measures are used in a continuous treatment design. The problem of carryover effects that interact with subsequent treatments is quite different. No answers to this problem are given here; instead we ask if there is, in fact, any way of preserving the advantages of a cross-over design and obtaining unbiased estimates of the treatment effects when carry-over interaction is present.

Let us take a hypothetical psychiatric experiment with a repeated treatment design. Say that a psychiatrist thinks slow reaction times are characteristic of paranoid schizophrenics and he wishes to alleviate this symptom by chronic administration of some tranquillizing drug. He selects a sample of N paranoid schizophrenics, puts each patient on a

*The word subject is used here as a general synonym for the experimental unit of observation.

maintenance dose and starts testing reaction time once a week. At the end of k weeks, the reaction time scores can be arranged as a rectangle, N rows by k columns. The statistical analysis indicated by such tests as Edwards (1950), Lindquist (1953) and McNemar (1949), would be a two-way analysis of variance, with k-1 degrees of freedom for the effect of weeks, N-1 degrees of freedom for the between-subjects effect, and (k-1)(N-1) degrees of freedom for the subject-by-week interaction effect. Then the significance of the differences between the k weekly means would be assessed by an F ratio using the subject-by-week interaction as the error term. Let us call this ratio the "univariate F."

One of the basic assumptions for the use of subject-by-week interaction as the error term, is that all ovserved scores are statistically independent of one another. However, in this hypothetical experiment, it is almost certain that the scores on the first week will have a positive correlation with the scores on the second week, third week, etc.

In 1948, Kogan suggested that if the assumption of independence is not met, the univaraite F ratio overestimates the significance of the difference between the k-means. In 1954, G.E.P. Box, in a brilliant article, gave a general technique assessing the effect of departures, from independence and from equal variances, on the univariate F. In general, his conclusions substantiate Kogan's guess; when the null hypothesis is correct and the observations are dependent, the univariate F will exceed the tabled significance levels more often than it should. Roughly speaking - the effect of correlation between the weeks (i.e., treatments) is to reduce the apparent number of degrees of freedom in the numerator and denominator of the F ratio.

Box's model, and the conclusions he drew, are worth sketching here since they demonstrate why multivariate analysis of variance, rather than univariate analysis of variance is most generally appropriate for correlated observations. Two assumptions are made:

a) The vector of scores for any subject is statistically independent of the score vector for any other subject, under the null hypothesis.

b) Each vector is a sample from the same multivariate normal population.

In terms of our hypothetical psychiatric study, this means that the N paranoids are randomly selected and the relation between the scores of any two weeks, say week s and week t, is bivariate normal. The variance of week t, v_{tt}, need not equal v_{ss}; v_{et} does not necessarily equal the correlation between any other pair of weeks.

C. R. R. Rao in 1952 (pp. 239-244) showed how Hotelling's T^2 could be adapted to give an exact test of the differences between correlated means. Basically, Rao takes a linear function of the k scores and

compares the mean of this linear function to the variance of the linear function. (A convenient computation routine for this test is given by T. W. Anderson in his 1958 text par. 5.3.5).

Using an exact multivariate approach, Box shows that, under the null hypothesis, the true distribution of the univaraite F with (k-1) over (k-1)(N-1)d.f. can be approximately represented by the same F value with the degrees of freedom reduced by a fraction, ε. This fraction, epsilon, is a function of the k by k covariance matrix.

$$(1)\quad \varepsilon = k^2(\bar{v}_{tt} - \bar{v}_{..})^2/(k-1)\left[\sum_{t=1}^{k}\sum_{s=1}^{k} v_{ts}^2 - 2k\sum_{t=1}^{k}\bar{v}_{t.}^2 + k^2\bar{v}_{..}^2\right]$$

where v_{ts} is the covariance of the N pairs of scores from week t and week s, and $\bar{v}_{tt}$ is the average variance for the k weeks.

The maximum value of epsilon is one, and this is reached only when the k variances are equal and the $\frac{k(k-1)}{2}$ correlations are constant. In this case, Box's approximation gives the exact results; when the correlations are constant and the variances are equal, then the univariate F ratio can be used to give the exact significance level of the differences between the k correlated means.

Geisser and Greenhouse (1958) have shown that the lowest value that epsilon can take is 1/(k-1). They argue that since no one has shown what sample estimate of epsilon is most appropriate, and the robustness of epsilon has not been investigated, it is best to use the minimum value of epsilon for a conservative test. This conservative test consists of computing the univariate F, and entering the tabulated F distribution with 1 over N-1 d.f. If the result is significant, there is no need to go further; the exact test would be significant. However, if the conservative test is not significant, one can now make an upper-limit test of the univariate F (setting epsilon equal to unity). If an assumed epsilon value of unity gives a non-significant result, then the null hypothesis can be accepted, since no calculated value of epsilon can give a more significant result. However, if using the full degrees of freedom gives a significant result, then the research worker is in a dilemma. Geisser and Greenhouse apparently would next try Box's approximate test, using a sample estimate of epsilon. I would recommend an exact multivariate test such as Rao's.

You can see that the Geisser-Greenhouse approach allows one to bracket the significance level of F with the same amount of computation that is used in the usual two-way analysis of variance. The laborious computations for an exact multivariate A. of V. include the data necessary for a two-way A. of V. Therefore, it will always be profitable to try the Geisser-Greenhouse approach first, before proceeding to the rest of the distasteful arithmetic necessary for multivaraate analysis.

Here it is essential to stop and point out that Box's model explicitly assumes multivariate normality. What alternatives do we have if multivariate normality does not hold or can not be forced by a transformation? As we mentioned previously, the Rao exact multivariate test for differences between correlated means essentially compares the mean of a linear function to the variance of that linear function. The question of multivariate normality can therefore be posed as the question of whether the scores produced by the linear function have a normal distribution. When k is large and correlations are near-zero, we know that the linear function will yield a near-normal distribution of scores. However, if the linear function scores are not normally distributed, the means will have a near-normal shape, assuming the samples of N subjects to be large and selected at random. Therefore the Rao multivariate test will be robust to deviations from normality when N is large or when k is large and the correlations are small.

In those cases where robustness is in question because of small N, high correlation, or other characteristics of the data it seems to me that the basic strategy should be to resort to the randomisation test introduced by R. A. Fisher (1935, par. 21). If we use Box's first assumption, that each subject's vector of scores is independent, and change Box's second assumption to read "each vector is a sample from the same symmetric multivariate distribution" then we will meet Fisher's requirement that the scores for the treatments be drawn from the same population. Since the problem is whether the means differ significantly, it seems reasonable to use the usual univariate "between treatment means" deviance as the criterion. However, E. S. Pearson (1937) has pointed out that the most powerful criterion depends upon the form of population distribution. For example, when the population distribution is rectangular, midpoints rather than means should be used. The null hypothesis here is that the k scores for any subject are completely interchangeable and any permutation of the k scores can be substituted for the original vector. Since there are N subjects there are $(k!)^N$ sets of scores. Each set is a possible sample from the original finite set of scores. The between-treatments deviance can be computed for each permutation and we can ascertain where our observed between-treatment deviance falls in the frequency distribution of all possible values from this finite sample. If our observed sample value equals or exceeds the assigned significance level, the means can be judged to be significantly different.

This permutation test preserves one of the advantages of the univariate A. of V. approach, N can be less than k. (The multivariate methods cannot be applied routinely for N less than k since the inverse of the k by k covariance matrix does not exist). One disadvantage of the permutation test for differences between means is the requirement that all treatments have identical distribution moments (except fot the means). However, the identical distribution assumption apparently is made in every parametric or non-parametric statistical test, of the difference between two or more samples. The assumption of identical distributions seems to be necessary for generating any statistical test of differences. Some empirical results I have seen suggest that if the distributions are symmetric about their

midpoints, they need not be identical; the permutation test is presumably robust to non-identical distributions in these cases.

The basic disadvantage of the permutation test is the extraordinary amount of labor required for even moderate values of N and k.

Suppose, instead of asking if the means are different, we ask if the scores for one week tend to be higher than the scores for other weeks. Then the hypothesis concerns the equality of the rank order averages.

As is well known, Kendall's W, or concordance coefficient, is a simple easily-computed test of this hypothesis. (1948).

Wallis and Friedman independently, and about the same time as Kendall, devised statistics that are algebraically equivalent to Kendall's W.

Essentially, Kendall's W is a permutation test on scores that have been transformed into rankings. The basic assumptions are - score vector independence and identical treatment distributions, exactly the same as those made for Fisher's randomization test, but the laborious computations have disappeared. However, it should be noted that we are now asking a different question - whether the average rank differs significantly between treatments. Does inequality of the average rank imply inequality of the means and vice versa? I have found several empirical examples where Kendall's W was significant but the univariate and multivariate A. of V. tests fell below significance.

Generally, one assumes that the rank order statistic and the A. of V. statistic are testing the same thing, but that the rank-order test is less powerful. However, the discovery of empirical examples where Kendall's W was significant and the F ratio wasn't, shook my faith in this proposition. Since then, I have learned how to construct examples where the means are exactly identical but the average rank differs significantly. However, in the construction of these counter-examples, I found it necessary to introduce non-identical distributions, to violate one of the two basic assumptions.

Therefore, I would like to raise the explicit question: What are the necessary and sufficient conditions such that rank-order tests are less powerful versions of the analogous A. of V. tests? This problem transcends the context of repeated measurements. Perhaps situations can be devised such that any rank-order statistic will be more significant than its metric analog. I raise this question - I hope some statistician can answer it.

I am saying that sometimes rank-order tests answer a different question than their metric analogs do. I am not saying that rank-order tests should be abandoned. There may be many occasions when the A. of V. test is not quite the right way to answer the question - when the major interest is in whether one treatment differs from another treatment, and the amount of the difference is irrelevant. There are other situations where it is not clear that the units of measurement are all equal,

as in psychological test scores, so that equal metric differences may not be of equal importance. In these and other cases, the experimenter, upon reflection, may discover that he is more interested in rank-order than in metric differences.

Let us now come back to our psychiatric example. You will recall that in our example the psychiatrist had placed his schizophrenic patients on a tranquilizer in the hope that the reaction times would be shortened. Time is a natural unit of measurement and there is little ambiguity there. If he is primarily interested in the therapeutic value of the drug, then the exact amount of decrease is important. Presumably, any improvement which is insignificant for practical purposes, say a decrease of 1/100 of a second, would be of little therapeutic interest, even if it were statistically significant. However, if his interest is primarily theoretical, for example, he hopes to find whether the delay is at the nerve-muscle junction or is caused by central factors, then any decrease in reaction time will be of interest to him.

Even if he knows that relative and not absolute differences are his main interest, should the psychiatrist use a general test of differences such as Kendall's W, or a test which specifies an a priori rank-order; for decrease in reaction time should be a monotonic function of number of weeks on the drug. Whenever a set of correlated means has a predicted rank-order, each subject's obtained rank-order can be correlated with the predicted rank-order and the average of all N rank-order correlations can be tested for significance. In 1954 Jonckheere presented an explicit test of this sort, using Kendall's tau. Lyerly (1952) has discussed the distribution of the average Spearman rank-order coefficient, rho.

Jonckheere's average tau test (as well as the equivalent Spearman form) is unique among non-parametric tests in that there is no parametric analog. So far as I know, there is no regression procedure or Hotelling T^2 criterion that can be applied to test for monitonicity. Any metric technique needs a formal specification of the exact mathematical relation between reaction time and weeks, before such a relationship can be tested.

This brief survey of the statistical tests appropriate to a continuous treatment design does not, of course, cover all the relevant topics, but it does show there are rational procedures for treating the data which differ considerably from those found in many statistical text-books.

So, to summarize the statistical recommendations in our hypothetical experiment, the psychiatrist might use the Geisser-Greenhouse multivariate A. of V. approach or he might use Jonckheere's average rank-order coefficient, but he should not make a routine application of the usual two-way A. of V.

Let me deal briefly with some of the experimental problems raised by repeated measurements. Almost certainly there will be an improvement in reaction time, whether or not the drug is used. The very act of measuring reaction time gives the patient practice on this task, allows him to adjust

to the situation, and so on. This quasi-Heisenberg effect is very common with most kinds of repeated measurements. The blood pressure of a subject is usually higher during the first few determinations than on subsequent occasions. The prick of the hypodermic needle can cause significant changes in blood composition until the subject becomes habituated.

One common way of dealing with the problem is to run a control group. This allows us to estimate the trend, without the drug. Another way is to run each patient through the measurement procedure until he reaches a steady state. Control groups are, of course, almost always necessary because of vagaries in the experimental situation, apparatus, etc., but even when controls are used, I advocate running each subject to a steady state. Not only do you eliminate any complex trend that may exist, but the intra-subject variation usually decreases markedly. This makes it particularly advantageous to use the intra-subject rather than the inter-subject variance as error.

But this raises the question of what part of the performance we want to measure. Perhaps it is exactly the factor in learning, habituation, practice, etc., which the experimenter wants to study. In this case, a control group will enable him to assess the effect of a drug on the initial rate of change. In most situations we are interested in the performance of the Subject on a well-learned routine task. When this is, in fact, true, then we may be measuring some factor which is irrelevant to our question when we include measurements taken at a time of rapid learning or habituation.

Let me hasten now to my final point, a sweeping generalized warning against the use of crossover designs.

If you wish to assess the separate effect of two or more treatments, don't apply the treatment to the same organization. A brief logical justification is as follows: if you're trying to assess the effect of a treatment by itself, then almost certainly you do not have enough previous data to estimate the carryover effect and in particular the interaction of the carryover effect with other treatments. But all designs using two or more treatments on the same organism assume that there is no interaction of the carryover effect with preceding or subsequent treatments.

Another way of looking at it is to consider the rotation experiment. Here the treatments are applied in predetermined sequence and the problem is the effect of the sequence of treatments on the subject rather than the effects of the individual treatment.

There are countless examples in medicine where the order is all-important, e.g., when weak and strong bacterial strains are injected in an organism. The enormous difference in the effect of the two rank-orders is the basis for vaccination.

If the experimenter who proposes to use a cross-over design thinks that a rotation experiment with the same treatments would also yield important information, he is assuming that carryover interaction can exist; that treatment A can inhibit or potentiate treatment B. In this case, his estimates of the effect of each treatment from the cross-over design will be hopelesssly enmeshed with the carryover interaction effects.

REFERENCES

1. Anderson, T. W. (1958). Introduction to multivariate statistical analysis. New York, Wiley.

2. Box, G. E. P. (1954). "Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and correlation between errors in the two-way classification." Ann. Math. Statist., 25, 484-498.

3. Edwards, A. L. (1944) Statistical Analysis. New York, Rinehart.

4. Fisher, R. A. (1935) Statistical methods for research workers. New York, Stechart.

5. Friedman, M. (1937). "The use of ranks to avoid the assumption of morality implicit in the analysis of variance." J. Am. Statist. Assn. 32, 675.

6. Geisser, S. and Greenhouse, S. W. (1958). "An extension of Box's results on the use of the F distribution in multivariate analysis." Ann. Math. Statist., 29, 885-891.

7. Geisser, S. and Greenhouse, S. W. (1959). "On methods in the analysis of profile data." Psychometrika, 24, 95-112.

8. Jonckheere, A. R. (1954a). "A distribution-free k-sample test against ordered alternatives." Biometrika, 41, 133-145.

9. Jonckheere, A. R. (1954b). "A test of significance for the relations between m rankings and k ranked categories." Brit. J. Statist. Psychol., 7, 93-100.

10. Kendall, M. G. (1938). "A new measure of rank correlation." Biometrika, 30, 81.

11. Kendall, M. G. (1948). "Rank correlation method." London, Charles Griffin & Co., Ltd.

12. Kogan, L. S. (1948). "Analysis of variance - repeated measurements." Psychol. Bull., 45, 131-143.

13. Lindquist, E. F. (1953). Design and analysis of experiments in psychology and education. New York, Houghton Mifflin.

14. Lyerly, S. B. (1952). "The average Spearman rank correlation coefficient." Psychometrika, 17, 421-428.

15. McNemar, Q. (1955). Psychological Statistics. New York, Wiley.

16. Pearson, E. S. (1937). "Some aspects of the problem of randomisation." Biometrika, 29, 53-64.

17. Rao, C. R. (1952). Advanced statistical methods in biometric research. New York, Wiley.

18. Wallis, W. A. (1939). "The correlation ratio for ranked data." J. Am. Statist. Assn., 34, 533.

THE GERMFREE LABORATORY AT THE WALTER REED ARMY INSTITUTE OF RESEARCH: Design of Experiments using Germfree Animals.

Ole J. Malm
Stanley M. Levenson
Captain Richard E. Horowitz
Departments of Germfree Research and Surgical Metabolism and Physiology
Walter Reed Army Institute of Research, WRAMC

Germfree rats, mice, guinea pigs and chicks are now routinely available in special laboratories like the Walter Reed Department of Germfree Research. The germfree animal has become a research tool, uniquely suited to provide answers which cannot be obtained by the use of conventional animals alone.

By the use of germfree animals, certain problems can be readily and equivocally answered in simple experiments which do not involve large numbers of animals and statistical evaluation of the experimental data. A fundamental question, asked by Louis Pasteur (1885) was whether life without bacteria was possible. This question has been answered in the affirmative by the successful rearing of a number of animal species over long periods of time by the pioneer laboratories in germfree research, (goat, rabbit, monkey, rat, mouse, guinea pig, fowl and fish).

Many metabolic processes occurring in the animal organism may be dependent upon enzyme systems of commensal bacteria rather than on endogenous enzymes in the animal. The germfree animal lends itself superbly for the study of these problems. It is possible, through a few well designed experiments, to obtain definite answers to a problem which requires a great number of complicated experiments when undertaken with conventional animals as exemplified in the following study of urea metabolism accomplished at the WRAIR Germfree Laboratory (1). The metabolism of urea, the first organic compound to be synthesized (Wohler, 1828), has always interested biologists and physicians. Considerable time and effort has been expended by large numbers of investigators in laboratories all over the world attempting to determine whether the metabolism of urea in mammals was under endogenous or bacterial control. In a review of this problem published in Physiologic Reviews, Kornberg listed over 50 investigations, yet the precise role of the intestinal bacterial flora remained equivocal and inferential. Indeed, as recently as 1956, Conway, a leading Irish biochemist, presented evidence before the 20th International Physiological Congress, which he interpreted as showing that the gastric urease of mice was intracellular rather than bacterial.

The problem of the bacterial origin of urease was clearly susceptible to test in the germfree animals. Accordingly the metabolic unit of the Germfree Laboratory, WRAIR, injected subcutaneously C^{14} urea into two conventional and three germfree rats, and administered it orally to one germfree rat. Each rat was then immediately placed in a metabolic apparatus and its urine, stools, and expired air were collected. Any hydrolysis of urea to ammonia and carbon dioxide would be readily detectable, since the CO_2 formed from the administered urea would contain radioactive carbon.

The conventional animal's expired air contained 100 times as much radioactivity as the germfree animal's. The pattern of urea hydrolysis in the germfree rats was the same whether the urea was given subcutaneously or intragastrically.

The very small fraction of the injected C^{14} (0.02%) expired by the germfree rat is due to spontaneous hydrolysis of urea, not to enzymic breakdown.

These results, conclusively, demonstrate that the enzymic hydrolysis of urea by the rat is effected only by the urease of its bacteria. Moreover these results provide the experimental answer to the clinical observation that certain oral antibiotics effectively control ammonia toxicity of patients with liver dysfunction. With a few germfree animals and in a very short period of time, an unequivocal answer to this problem which had been inconclusively worked on by many investigators for over 75 years was obtained.

Unfortunately, many experiments in which germfree animals can be of singular value, involve a more complicated design due to some special problems in germfree research. These special problems fall into two main categories:

1. The special environment in which the germfree animal lives, and

2. Peculiarities inherent in the germfree animal itself.

In the discussion to follow we will define some of these environmental and biological factors peculiar to germfree research. The main problem is to devise the proper control for the germfree animal when the control is to be his normal or conventional laboratory counterpart. This is a vital question since a well controlled experiment, properly planned, will save time, work and money by reducing the number of animals necessary to obtain statistically significant results and obviate repetitions.

THE "GERMFREE" ENVIRONMENT. The germfree environment is potentially the most controllable of any now available in which to conduct animal research. Ideally, in any experimental study, the investigator would strive at following "the dictum of the single variable." In order to do so, he must know his experimental system, including the environmental conditions of his animals, in every detail and duplicate the conditions to which the experimental animals are subjected as closely as humanly possible in the controls.

Diet, temperature, humidity, ventilation, illumination, caging, noise, handling and gentling of animals are factors which should be under continuous control in acute as well as chronic type experiments. One must have constancy of the exterior milieu so as not to disturb the homeostasis of the internal milieu, except by the experimental variable under study.

We do not know to what extent minor and uncontrolled variations in one or more of the environmental factors mentioned, may influence the performance of animals in a given experiment. It is because of this lack of specific information on several counts that one should control all known

variables in the experiment. Otherwise, differences found between experimentals and controls may be ascribed to the experimental variable, while in reality the observed difference was mainly due to uncontrolled variations in one or more environmental factors.

Let us first consider housing and caging of germfree animals. The Reyniers type steel tanks (Figures 1, 2 and 3) used in our laboratory provide protection of the animals to air-contamination through a filter system in the inlet air and a germicidal trap for the outlet air. However, there is a rather brisk and steady flow of air (5 cfm) under slight positive pressure, which affects temperature, humidity and barometric conditions in the tank. Furthermore entry into the tank is limited to the glove ports and the autoclave route. The animals can thus only be handled by hands protected by thick rubber gloves plus cotton work gloves. The handling and fondling aspects and their possible influence upon the reactions and emotions of the animal are largely unknown as experimental parameters. We should recognize this fact and equalize conditions whenever possible.

With regard to caging, the limited space in each tank might tempt the investigator to use small restraining cages, and even to cram two or more animals into each cage. This is of course only permissible if controls are housed in an identical way, although there is usually no need for such extreme space economy in our animal rooms.

In many experiments, especially where influences of dietary factors are under study in germfree versus conventional animals, the temptation to house more than one animal in a cage should be overcome. If one animal dies and is cannibalized by the survivor, the experiment may be ruined. If the cage of the germfree animal is of a type which limits coprophagia, the cage of the control animal should be identical. The feces eaten by the conventional animal are not the same as those eaten by the germfree. The conventional feces contain bacterial body constituents, but even more important, vitamins synthetized by the bacteria of which the vitamin B-group may be the most important.

With regard to temperature inside the germfree tank, this is a function of seven factors: The temperature of the inlet air, the rate of air flow and the humidity, the temperature of the room in which the tank is located due to ready convection of room temperature through the steel walls, the illuminating lights, the animals own heat production and last, but not least, to heating incident to operation of the autoclave attached to the tank when entry or exit of material is necessary.

The tank temperature can be controlled within rather narrow limits by special devices; the point is that temperature variations induced in the germfree tank should be duplicated for the control animals at the same time. The marked influences of environmental temperature on a great number of biological phenomena are well known and need not be detailed here. It suffices to mention as examples (2) that growth rates, dietary requirements, physical activity, sexual cycles and functions, mitotic activity and renewal rate of the epidermis are all markedly influenced by the environmental temperature. Environmental temperature also affects survival rates following different types of trauma, like hemorrhage shock, tourniquet shock and burns. (3)

Although the sensitivity of animal functions is not as pronounced to changes in humidity as in temperature, major and uncontrolled variations should be avoided. The requirements for optimal levels of humidity, as for temperature, vary with age and species of animals. Temperature and humidity affect energy exchange in all warm-blooded animals. Particularly in the stressed animal and perhaps especially in burn studies, humidity and temperature control are mandatory. (4)

The illumination requirements of animals cannot be accurately defined today. A constant day-night cycle seems to be particularly important for rodents. Thus seasonal variation in breeding can be reduced or eliminated. (5) Illumination for paired experiments must be of the same intensity and wave length for it is known that light of different wave lengths has profound influences on adrenal functions. (6)

Noise as a potentially important factor is not well understood in its disturbing effect on animals in a secluded environment like the steel tank. All we can do, is to equalize the noise factor by the simple rule: if you bang the experimental tank A, bang tank B, housing the controls.

Diet is another very important factor needing control due to the special processing needed for germfree animals. It is evident that equal conditions for experimental animals and controls imply that both get the same diet. The diet for germfree animals is autoclaved prior to entry, and when distributed in the tank, it is not subject to attack and alteration by bacterial contamination. Not so for the conventional animals. Even if the diet is autoclaved under the same conditions as for the germfree animals, the similarity may end here. As soon as the diet is cooled and distributed to the conventional animals contamination with its manifold implications will take place. We do not know how to completely equalize the factors influencing the diet in experiments involving both germfree and conventional animals. To illustrate our attempt towards this end, some details from a current series of long-term experiments carried out in collaboration with NIAMD on germfree and conventional rats on a choline-deficient, cirrhosis producing diet will be briefly summarized.

The diet is mad up identically by the same person for three groups of rats, (1) germfree in sterile tanks, (2) conventional rats in nonsterile tanks, and (3) conventional rats in our ordinary rodent room. Ingredients, weighing and mixing, and sterilization procedures are identical. The water supply is identical for all groups, only canned U. S. Coast Guard Emergency water is used. Food is offered in equal amounts to all groups on Mondays, Wednesdays and Fridays.

It is evident that identical environmental conditions for germfree experimental animals and their conventional controls, apart from the presence of bacteria in the environment of the latter, necessitate that the controls are also kept in tanks in the same room. Air flow rate, pressure, temperature, humidity, illumination, handling and noise can thus with proper care be canceled out as experimental variables. A third group of animals, conventionals in ordinary animal rooms, should ideally be set up to distinguish differences in reaction to an experimental variable between conventional

controls housed inside tanks versus controls in the animal rooms. Only by careful analysis of such triple-phased experiments can we learn more about the relative importance of the environmental factors discussed previously.

Now to the germfree animal itself. The main know physiological differences between the germfree and the conventional animal involve the cellular and humoral defense mechanisms, especially the roticuloendothelial system (RES), and as a corollary, certain of the plasma proteins; also the gut, especially the cecum of the rat and the guinea pig. (7,8,9)

1. THE STATE OF THE CELLULAR AND HUMORAL DEFENSE MECHANISMS IN THE GERMFREE ANIMAL. By definition, the germfree animal is free of demonstrable bacterial and fungal infections by the culture techniques used to establish germfreeness. The animal does not harbor parasites, as determined by fecal screening for eggs and parasites and careful autopsy. While most workers probably feel that exogenous viruses are not present in germfree animals, the situation is not clear with regard to viruses which may be transferred to the fetus in utero or (possibly) through the milk in suckling rats and mice born of germfree parents. This unsettled status of the germfree animal with regard to viruses is unfortunate if germfree animals are used in experiments designed to study development of tumors in cancer research, and of course, in experiments with viral agents in a presumably virgin organism. The absence of a live micro flora accounts for the unstimulated state of the lymphoid tissue and particularly for the low numbers of plasma cells seen in the tissues of the entire gut of germfree rodents and birds.

It is, however, important to realize that the germfree animal is exposed to antigenic challenge by foreign materials and that while his RES is underdeveloped anatomically and possibly functionally, it is certainly not dormant. Bacteria, and maybe viruses, are always present in the diet when prepared. Infectious agents are killed by autoclaving, but lipopolysaccharides and heat-coagulated bacterial proteins may enter the germfree organism and act as antigens. Protein material from the food itself is another source of antigens. While the supply of bacterial antigenic material must be substantially less in the gut of the germfree animal, the situation is not different with regard to antigens offered with the food itself. The underdevelopment of the RES refers particularly to the lack, or scarceness of, nodular lymphoid structures in the gut, while "free" or scattered RES elements, including plasma cells, are always found in the mucosa and submucosa to an extent of 10 to about 30 per cent of that seen in conventional animals of the same species. The status of the RES elements in the respiratory tract of the germfree animals remaind to be studied in detail.

The low intensity of challenge by RES-stimulating antigens must be kept in mind in the design of, and especially in the interpretation of, experiments involving traumatic procedures like hemorrhage, traumatic shock, radiation injury and burns. At our present state of knowledge, it is naive to interpret differences in survival or tolerance to any one of these procedures between germfree and conventional animals solely to the presence or absence of bacteria. In any situation involving tissue injury, the germfree animal must presumably be in a different position to take care of the consequences of massive cellular destruction.

At the present time, the conventional animal serves as a control for his germfree counterpart, or vice versa, only if a marked difference in two variables is accepted and taken into account in the interpretation of experiments involving tissue injury:

a) the conventional animal contains bacteria and,

b) has a normally developed RES.

If in the future one could achieve a "normal" development, at least in terms of tissue mass, of the RES in the germfree animal by nonspecific stimulation with one or more injected or fed antigens, experiments involving tissue injury may become more meaningful with regard to the effects of the crucial variable - the presence or absence of bacteria.

Another project which, if successful, will enhance the usefulness of the germfree animal as a tool of research, is the production of a nutritionally complete, wholly synthetic diet which is hypo-allergenic or, ideally, non-antigenic. Several laboratories, including our own, are presently engaged in this work starting from the soluble, synthetic diet of Greenstein and Birnbaum. (10) This type of diet will permit basic studies of immunologic and defense mechanisms, including the physiology and biochemistry of the RES under completely controlled conditions.

THE PLASMA PROTEINS. The concentration of gamma globulins and the carbohydrate-rich alpha globulins are lower in germfree animals than in their conventional controls. These proteins are synthetized mainly or exclusively, by cells belonging to the RES. When the germfree animal is challenged with antigenic material, especially live bacteria, the RES is activated, and in some weeks the plasma protein spectrum cannot be distinguished by ordinary chemical and electrophoretic methods from that of a conventional animal. (11,12,13)

THE GUT AND CECUM. Smaller villi and very scant development of lymphatic structures are characteristic of the germfree state. The most striking difference, however, is the markedly increased cecal volume in germfree mammals. The cecum with contents weighs on the average 3 - 5 times as much in most germfree guinea pigs. The cecal wall structures seem underdeveloped, thinner, and the water content of the cecal contents is higher. This finding has tentatively been interpreted to indicate active transfer of water from the plasma to the cecal contents, since the water content of the lower ileum fluid entering the cecum, is less, and not different from that found in conventional animals on a similar type diet. The large cecum gives rise to a rather high incidence of fatal volvulus, especially in the guinea pig. The trapping of substantial amounts of total body water in the cecal fluid is a complication in all experiments which will induce shock, for example, hemorrhage, tourniquet, and burns. An added control in this type of experiment may be cecectomized animals. Such preparations have been made successfully at the Lobund Institute by Dr. Gordon and his associates.

By way of closing the discussion of the many factors which need control in germfree research, we will give a brief account of an experiment which may turn out to be crucial in clarifying the alleged role of bacterial endotoxins as the agents which may be responsible for so-called "irreversible" hemmorhagic shock.

In every war, shock has been the major emergency complication of the wounded solider, and this problem will be even greater in any future war. Considerable delays in the treatment of civilian and military casualties caused by thermonuclear warfare must be anticipated. "Irreversible" shock will become a clinical problem of a magnitude never before encountered. (Irreversibility is a state of refractoriness to treatment in which the best available treatment fails to prevent or only delays circulatory failure and death). During the past 25 years, circumstantial but impressive evidence has accumulated which suggests that while lessened blood volume is the primary cause of shock, the development of irreversibility after severe hemorrhagic or traumatic shock is due to the entry of bacterial endotoxins into the circulation. This hypothesis, championed by Fine and his group in Boston (14,15), states that severe hypoxia in the bled, hypotensive and shocked animal, will lead to a breakdown of the normal gut-blood barrier to bacteria and endotoxins and allow absorption or entry of bacterial endotoxins into the circulation. Endotoxins from gram-negative bacteria normally present in the intestinal flora, will, when introduced into the circulation, augment the already severe arterial hypotension by vasodilatory effects and result in collapse of the circulation, followed by death. Furthermore, the RES in the shocked animal has a markedly reduced phagocytic capacity towards potentially harmful macromolecules like bacterial lipopolysaccharides. Therefore, circulating endotoxin in amounts which in the non-shocked animal will be readily taken care of by the RES, is now free to exert its deletary effect in the shocked animal.

Obviously, this hypothesis could be put to test in the germfree animal. Such experiments have been reported by McNulty and Linares, (16) at Walter Reed and Zweifach, et al. (17) working at Lobund. Both groups used the germfree rat and found no significant differences in survival rates of germfree and conventional rats subjected to identical surgical procedures. In other words, germfree rats subjected to a bleed-out procedure and maintained at a fixed low level of arterial blood pressure for four hours, will upon retransfusion of the shed blood recover or die in numbers which are no different from that observed in the conventional rats. Taken at face value, the inference would be that bacteria on their endotoxins are not involved in the irreversibility of hemorrhagic shock and death in the germfree rat, which dies with the same gross and microscopic anatomical lesions found in the conventional animals. Hemorrhage into the small gut and injury to the mucosa, are characteristic features of the autopsy findings in irreversibly shocked rats. On the basis of these experiments in germfree rats, one cannot, however, discard the endotoxin-hypothesis as disproved. Small amounts of bacterial endotoxins, arising from heat-killed bacteria in the diet, are undoubtedly present in the germfree rat, some may be stored in the RES elements in the mesenteric lymph nodes. This endotoxin may be released during hypoxia, and additional small amounts may be absorbed from the intestinal contents during the hypotensive period and not be taken care of by the RES.

The argument is that these small amounts of circulating endotoxin are enough to precipitate irreversibility because the germfree animal with his anatomically underdeveloped RES is less resistant to endotoxin.

Experiments by Dr. Einheber in our laboratory with injection of a purified E. coli endotoxin which in a sufficient dose will kill the normal and the germfree mouse and in lesser doses induce a period of prostration and hypotension, showed, however, that this germfree animal is no more sensitive to endotoxin than his conventional control. The matter rests here at the present time. Definitive experiments to test the hypothesis must await production of a truly endotoxin-free, germfree animal, maintained on the synthetic hypo-allergenic diet, with and without an artificially stimulated RES.

SUMMARY. The design problems inherent in research with germfree animals have been described, specifically in regard to peculiarities of the germfree environment and animals. Methods for control of environmental and physiological peculiarities which permit investigators to follow "the dictum of the single variable" are discussed.

Planche 1

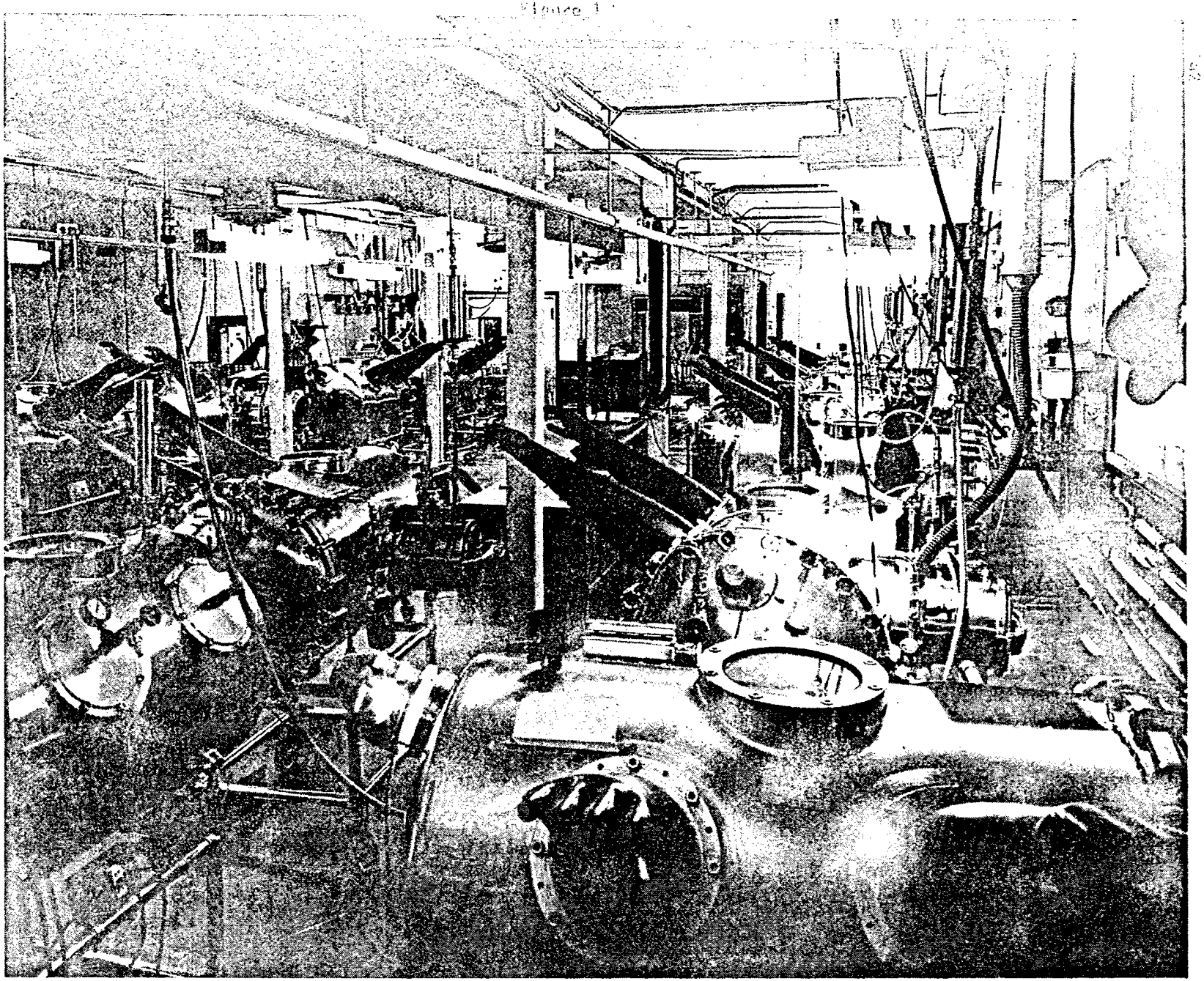


Figure 1

The tank room of the Department of Germfree Research at the Walter Reed Army Institute of Research.

Figure 2

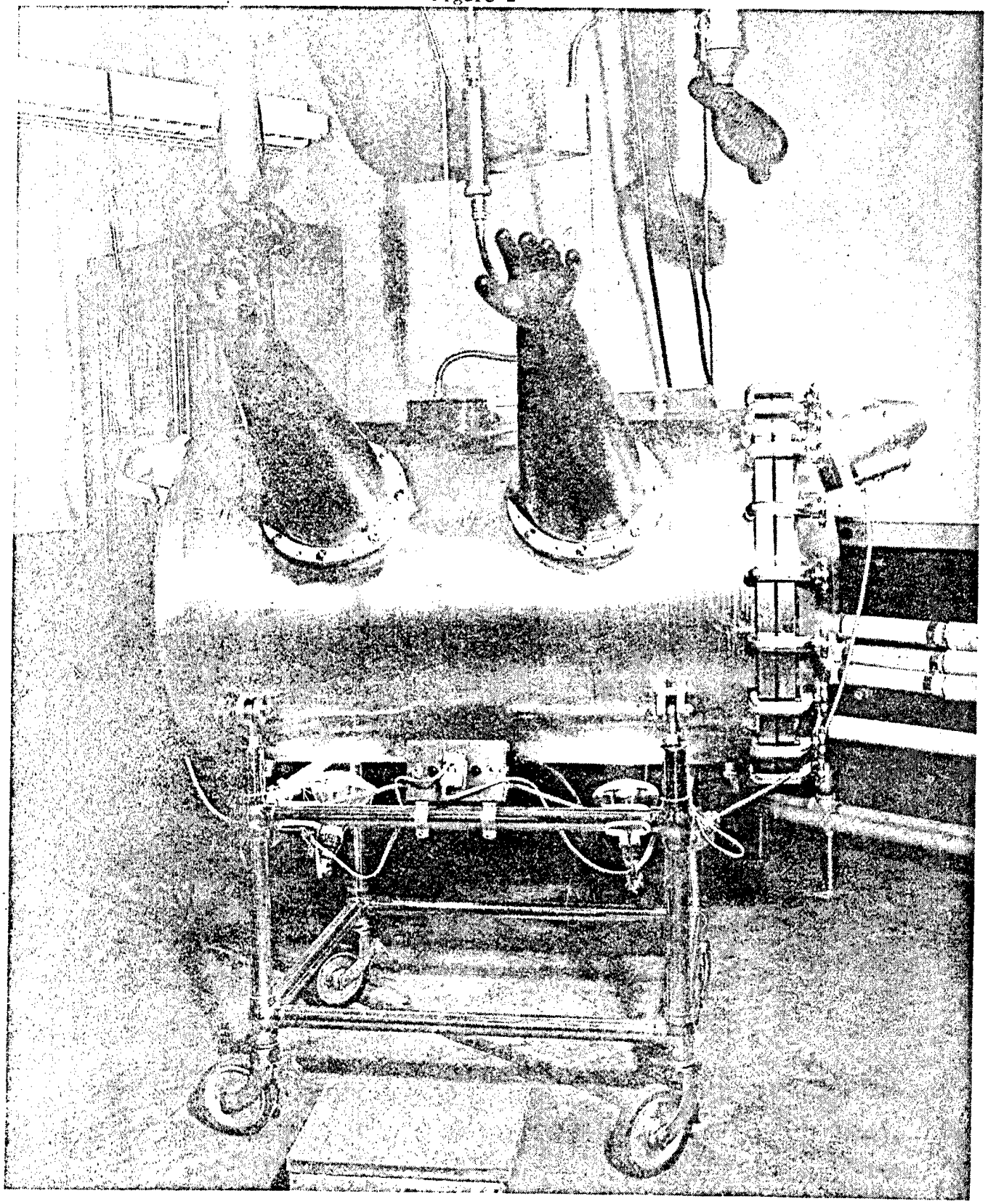


Figure 2

A Reyniers type heavy stainless steel germfree tank. The Department of Germfree Research at the Walter Reed Army Institute of Research uses 16 such one-man tanks as well as 4 similar tanks designed for two-man operation.

Figure 3

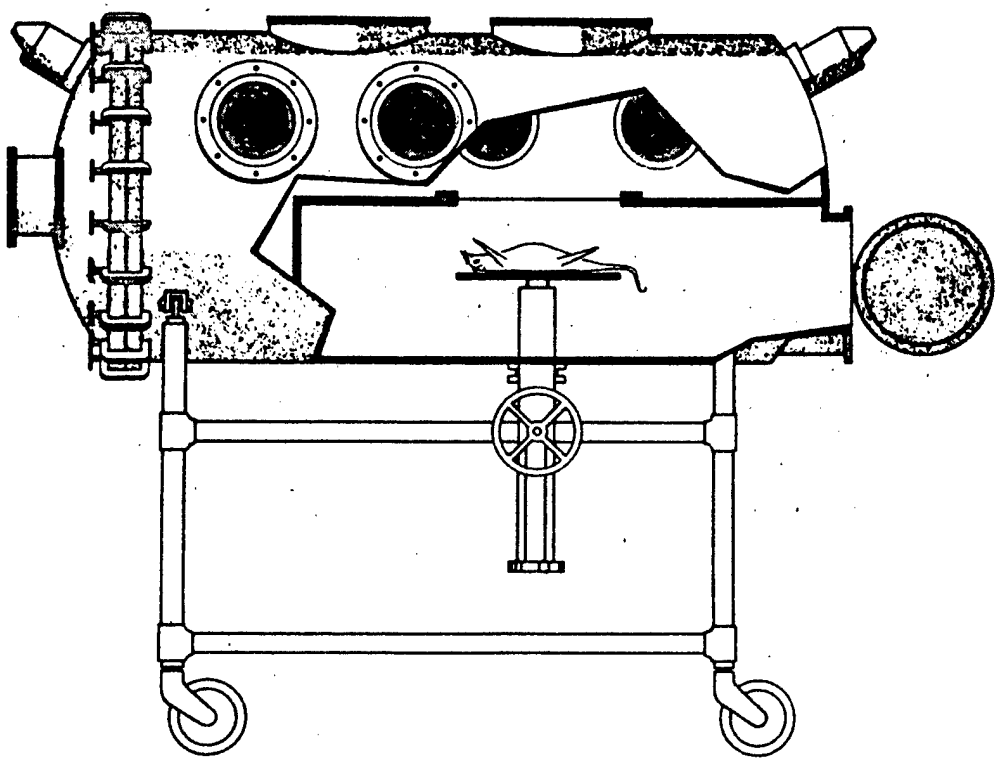


Figure 3

Diagrammatic cross-section of the germfree operating unit. The pregnant animal is introduced through the lower door into the lower section and is placed on the operating table. An elevator raises the table to the opening, which is covered by a sterile plastic sheet. Operator's hands are introduced through the glove ports while the view ports permit observation of the operative field. Caesarean section is performed by incision with a cautery through the plastic sheet, the heat of the cautery fusing the plastic to the skin. The uterus containing the young is excised and pulled into the sterile upper section of the tank, where the young are removed from the uterus, stimulated and breathing provoked.

REFERENCES

1. Levenson, S. M., Crowley, L. V., Horowitz, R. E. and Malm, O. J. The metabolism of carbon-labeled urea in the germfree rat. J. Biol. Chem., 234:2061, 1959.

2. Mills, C. A. Influence of environmental temperatures on warm-blooded animals. Ann. N. Y. Acad. Sci. 46:97, 1945.

3. Rosenthal, S. M. Experimental chemotherapy of burns and shock. Publ. Health Rep., 57:1923, 1942.

4. Pence, D. and Lindsey, D. Effect of high environmental humidity on survival of mice following an experimental burn. Am. J. Physiol., 195:719, 1958.

5. Alexander, D. P. and Frazer, J. F. D. Interchangeability of diet and light in rat breeding. J. Physiol., 116:50, 1952.

6. Gillissen, G. Uberblick und Versuch einer morphologischen Objektivierung sensorischer Einflusse auf dem, Organismus Arch. Hyg. 137:335, 1953.

7. Gordon, H. A., Doll, J. P., and Westmann, B. S. Effects of the "normal" bacterial flora on various morphological characteristics of the animal host: A comparative study of germfree and normal stock chickens, rats and mice. Anat. Rec., 130:307, 1958.

8. Wostmann, B. S. Serum proteins in germfree vertebrates. Ann. N. Y. Acad. Sci., 78:254, 1959.

9. Thorbecke, G. J., Gordon, H. A., Wostmann, B. S., Wagner, M. and Reyniers, J. A. Lymphoid tissue and serum gamma globulin in young germfree chickens. J. Infect. Dis., 101:237, 1957.

10. Greenstein, J. P., Birnbaum, F. M., Winitz, M. and Otey, M. C. Quantitative nutritional studies with water-soluble, chemically defined diets. I-V. Arch. Biochem. Biophys., 72:396-456, 1957.

11. Wostmann, B. S. and Gordon, H. A. Changes in the serum protein pattern of germfree rats upon exposure to a conventional bacterial flora. IV. Internat. Congr. Biochem., Vienna, Sept. 1958.

12. Gustafsson, B. E. and Laurell, C. -B. Gamma globulins in germfree rats. J. Exp. Med., 108:251, 1958.

13. Gustafsson, B. E. and Laurell, C. -B. Gamma globulin production in germfree rats after bacterial contamination. J. Exp. Med., 110:675, 1959.

14. Fine, J., Frank, E. C., Ravin, H. A., Rutenberg, S. H. and Schweinburg, F. B. The bacterial factor in traumatic shock. New Engl. J. Med., 260: 214, 1959.

15. Fine, J., Rutenberg, S. and Schweinburg, F. B. The role of the reticuloendothelial system in hemorrhagic shock. J. Exp. Med., 110:547, 1959.

16. McNulty, W. P., Jr. and Linares, R. Hemorrhagic shock of germfree rats. Am. J. Physiol., 198:141, 1960.

17. Zweifach, B. W., et al. Irreversible hemorrhagic shock in germfree rats. J. Exp. Med., 107:437, 1958.

THE DEVELOPMENT OF PARAMETERS FOR DETERMINING THE RESISTANCE OF SELECTED MISSILES COMPCNENTS TO MICROBIOLOGICAL DETERIORATION

C. Bruce Lee
Physical Sciences Laboratory, Research and Engineering Directorate, OTAC

The recent development of the Military Missiles Program in this country has necessitated a re-evaluation of the procedures in practically all phases of microbiological research, development and testing. This fact has been brought about by the number and complexity of the new materials employed in missiles, the peculiar designs and engineering of the components, and the problems of storage and ultimate operational requirements.

Deterioration microbiologists engaged in military activities realize the importance of the preceding statements and have found it necessary to develop for missiles research new parameters for testing. Further, there has been a need to adapt and re-evaluate those already in use, and to undertake research in order to assure to the manufacturers of missiles and missiles components that microbiological deterioration, specifically fungus action, will not be a factor in the malfunction of missiles once they are operational.

Missiles, missiles systems and their components are unusual in that there is a unique interdependence of items upon each other and all materials incorporated into a system must be verified absolutely reliable if the missile is to be, and remain, a tactical item. Thus, assurances of reliability must be secured by undertaking microbiological aging and deterioration testing in order to assure that there will be no difficulties traceable to the deteriorating effects of fungus action in the manufacture, storage and operation of the items.

For those present who may be unaware of the national program on military microbiological deterioration, a brief recase since its inception in the early period of World War II will be given. With the outbreak of hostilities, and movement of conflict to the tropics of the world, the military establishment suddenly found itself confronted with a monstrous problem; biological in origin, in which fungi, minute plants, were actually ruining and rendering unserviceable millions of dollars worth of critical materials by a natural ability to utilize in their metabolism the substrates supplied in the composition of military materials.

These fungi, the minute plants, are incapable of performing photosynthesis and, thus, they differ from the large familiar green plants which can make their own food. The species of fungi that are of concern to the military are usually microscopic or barely macroscopic in detail and all must secure their nutrition from pre-formed sources. I imagine that there are many here this morning who have vivid memories of food and clothing spoilage during tours of duty in the tropical areas of the world. The majority of the deterioration fungi reproduce most commonly by spores. These are shed into the atmosphere, the soil or water and, being easily air borne, they come to rest on a host of materials. If the material is susceptible to fungus growth, growth will proceed from the substance of the material substrate which is the pre-formed food necessary for fungus metabolism.

Usually, hydrocarbons and various minerals are the most easily metabolized sources of nutrition. Whatever the available nutrition, however, fungus growth on materials partially or completely destroys the material. Growth may be surface, or it may proceed internally, with the fungi producing thread-like mycelium, the first indications of fungus growth and presence.

During the war, fungus action was reported on a large number of items ranging from optical instruments to textiles. Most important, the spectrum of materials available for attack was almost entirely natural-in-origin, and included items which were cellulosic, proteinaceous or possessed animal or vegetable fats in their compositions. Control, was perforce, expedient and necessitated the complete discard and replacement of affected components, or the application of crude, surface-applied fungicides which often ruined serviceability of items by altering the physical or chemical properties to such an extent that the concerned material was rendered useless for military applications.

As a result of these experiences, the government entered the field of microbiology and sponsored basic and applied research on the control of fungus attack of military materials with the result that over the years, numerous tests have been developed which are capable of laboratory application in specification procedures. Particular efforts have been made to include specific tests for specific items. The resulting specifications invariably designate certain strains of species of fungi which have proven superior ability in degrading particular types of materials on the basis of origin and composition. For example, reference to any fungus test specification will reveal the prescription for the use of a single or species in tandem, and which may be cellulolytic, proteolytic or lipidophylic in degrading ability.

During the war years the employment of various synthetics in military items was begun, and since the cessation of hostilities, this use has expanded until currently, the role of synthetics in military goods far exceeds the natural-in-origin products in many items. At the beginning of the government efforts in the control of fungus deterioration, there was scant concern with possible fungus utilization of ths synthetic products; it being assumed generally, that these were inherently resistant. However, it was not long before the first incorporate uses of synthetics into military items that testimonies from the services revealed the fallacies of this assumption. Evidence was presented that synthetics were often excellent metabolic sources of nutrition for fungi with resulting alterations in physical and chemical properties. Conferences by government microbiologists on this problem resulted in a common approach to the entire field of fungus attack on materials and resulted in the following conclusions for national use:

1. In instances where materials have been found susceptible to utilization by fungi, such should be withdrawn from use and substitution accomplished employing funginert materials. The wide diversity of presently-available synthetics makes this possible.

2. Superior design and engineering of items must be employed from the concept stage until final manufacture and should take advantage of primary and continuing advice and suggestions of the microbiologist in order to eliminate loci of possible fungus utilization in any part of the completed assembly. (See Illustration 1 on the following page.)

Basically, these two suggestions have been followed by military microbiologists and the most important and pressing problems have been mainly solved or controlled.

Aside from basic and applied research, the military microbiologist functions currently as a consultant in development and testing. Further, suggestions are made concerning materials use and advice is given on design of components to withstand microbiological attack. In instances where there is no inert substitute for a susceptible item, advice is rendered regarding the use of possible fungicides. The present list of these chemicals is immense compared to years past and many have been developed for particular needs and uses. Contemporary fungicides rely on incorporation or compounding into a product, as well as on surface application. They take advantage of chemical and physical properties with a minimum of alterations to an item's characteristics.

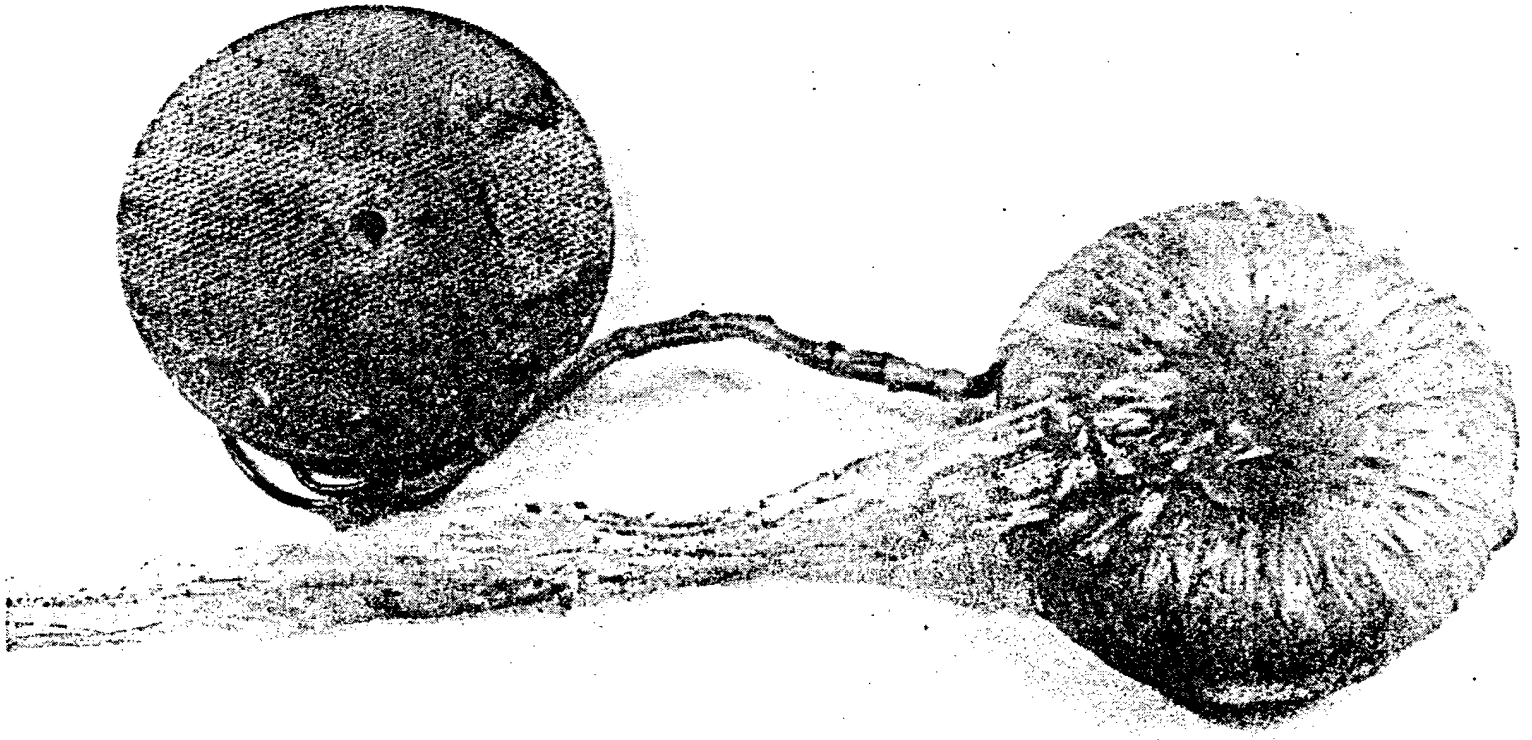
The requests of missiles manufacturers to our installation for information relative to their products' fungus resistance introduced a new phase of testing. Previous activities had been concerned with our mission for tank and tank-automotive vehicles and equipment. Usually, these materials were tested in part and a whole assembly was rarely submitted, although our facilities are geared to accommodate a six-by-six truck. Because of a strategic location in the automotive development center of Detroit, our organization was confronted suddenly with missiles measuring nearly sixty feet and with diameters from five to eight feet. The speaker is still amazed with the first request from a missile manufacturing service, "Can you expose this to fungus attack." <u>This</u>, being a missile nearly sixty feet long!!!!! (Illustration 2).

Missiles and missiles components submitted to fungus exposure were the Redstone, parts of the Jupiter and the entire Honest John. In addition, this organization had knowledge of research performed on the NIKE at a private installation.

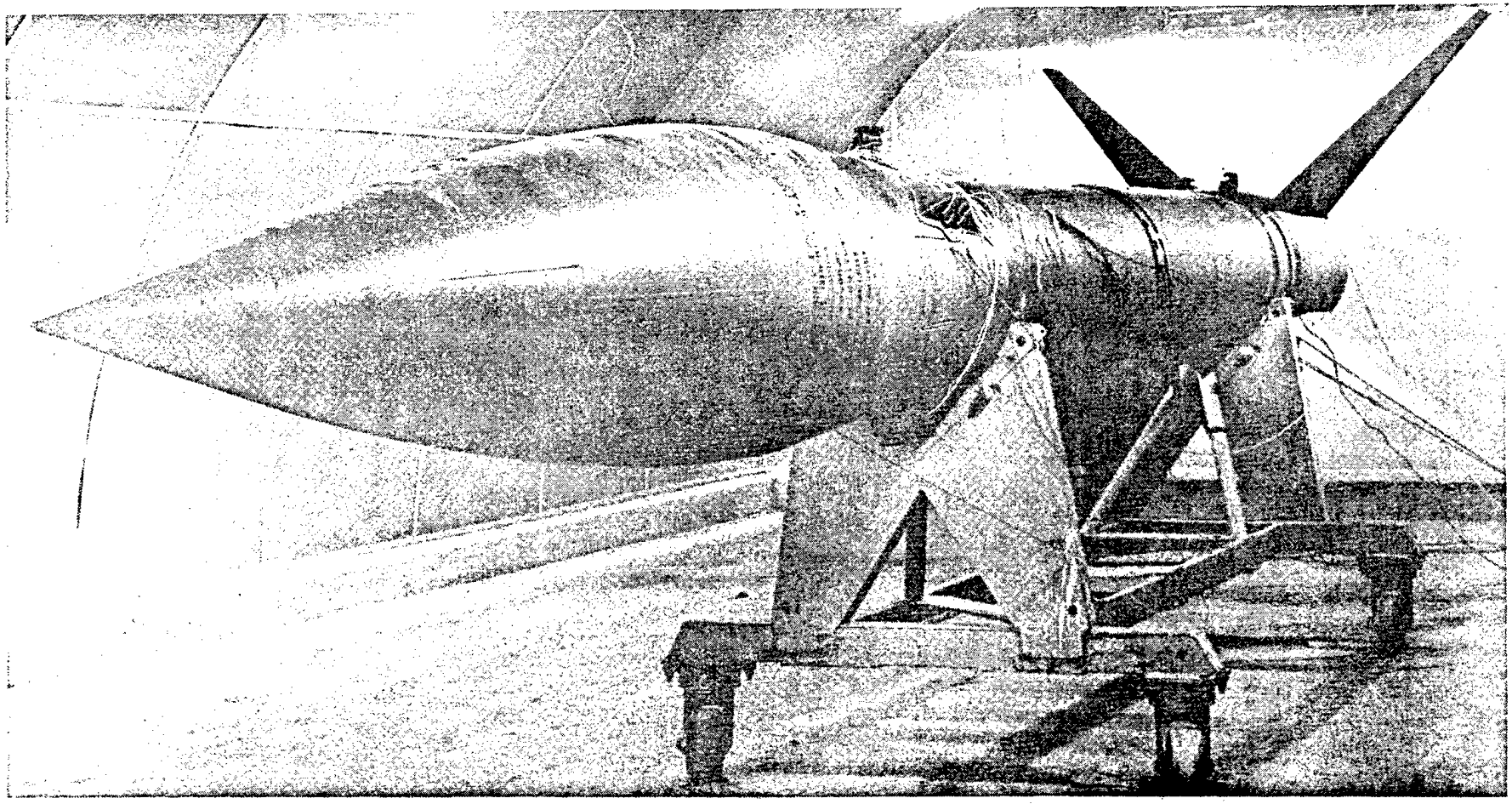
Literature surveys of the available specifications revealed no references specifically concerned with missiles, or the great variety of unique materials incorporated into these tactical weapons. Thus, our problem for the past few years was clearly indicated; the development of test parameters that would define the behavior and resistance of missiles and their components to fungus attack.

In addition to the whole or disarticulated missiles, information on fungus resistance was also desired for a large heterogeneous selection of materials, parts, partially-assembled systems and standard and specialty items. Many of these materials are synthetic in origin and prime in their use on military missiles components and were originally employed as it had

Illustration 1



T-2 transformer from the 1800 VA inverter from the Redstone Missile, CCMD. Silk winding removed in part to reveal the presence of fungi on the silk.



Honest John Rocket, Douglas, positioned within the Tropical Chamber at the White Sands Missile Range, New Mexico. Photograph included gives an idea of the size of missiles submitted for microbiological exposure.

been found that the natural-in-origin materials did not, and could not be expected to function adequately under the special conditions of temperature, humidity, etc., introduced by the storage and intended operation of the missiles.

To determine the procedures necessary for performing microbiological research on missiles, certain objectives were proposed for the investigations and these are:

1. What is the overall role of fungi in the utilization of missiles materials?

2. Is the missile item upon which fungi are growing actually being degraded and rendered unfit for service, or is the growth adventitious?

3. What items, or parts, are susceptible to fungus utilization and what are resistant?

4. What species of the lower fungi are to be indicated as materials degraders and are the materials susceptible to only one species, or is their utilization by several or more species?

5. What methods can be devised to demonstrate the effectiveness of protection and corrective measures necessary?

The narrative of the paper this morning will interpret the experience in the laboratory on missiles research in the light of the preceding points.

Further, in order to establish parameters for undertaking the research it was necessary first to arbitrarily limit the parts of the missiles which would concern the microbiologist. This was accomplished by an overall inspection at the site of manufacture. Some of the missiles tested were large and others easily accommodated into our testing facilities.

The inspection established the first parameter of testing; the fact that our research would be limited and conducted on the tail sections of the missiles. These are the parts containing the motor and control instrumentation, as well as the electrical connections. Samples of materials used in other parts of the missiles were requested and conclusions also submitted on their behavior to microbiological attack. Limitations in the size of the missile parts for testing were dictated by the accommodations available.

One of the important parameters in the fungus investigations was the choice of the testing situation. In previous experiences involving tank and tank-automotive components, all testing was accomplished in the laboratory and involved various pieces of environmental equipment. Owing to the size and diversity of materials used in missiles, decision was necessary as to testing site. Previous experiences of the speaker have indicated that wherever possible, it is more advantageous to employ the natural situation. According to location and program financing, various installations of the

country have secured data on the microbiological resistance of components in the field using such places as the tropical rain forest, the savannah, the desert or shore locations; places in which the temperatures and humidities are varying optimal for the development of the lower fungi in demonstrating their ability to degrade materials. When the missile research at Detroit first commenced, it was decided to press for running the research in the actual tropical rain forests available in Puerto Rico, or the Panama Canal zone. However, as a result of financial limitations on funding, the Detroit group was forced to confine work to the Detroit area and to arbitrarily choose a parameter of our own devising, the simulated tropical conditions afforded in the use of the tropical room.

The Detroit tropical room has been employed over a period of eight years for automotive testing and has been developed to attain conditions of humidity and temperature that are simulations of nature in offering optimal conditions for fungus development within the room and on materials placed into the room for evaluation. The room is a large structure, 20 feet long by 15 feet wide and with 9 foot ceilings and 8 foot access doors.

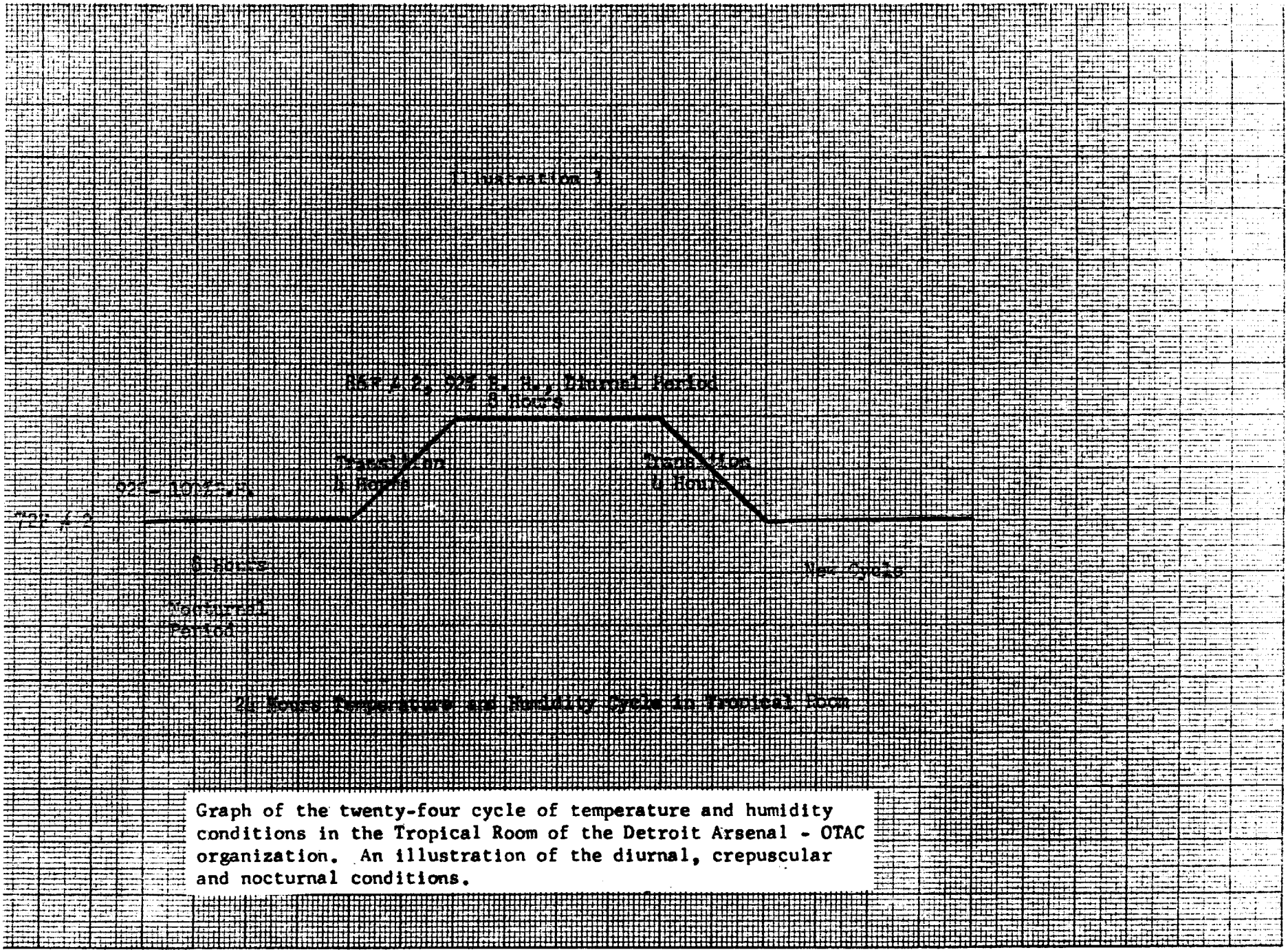
The conditions of temperature and humidity are original with the Detroit group and were determined on the basis of data available from the meteorological records of the rain forests of the world.

The simulation of conditions in the room, a phase of the parameter of the testing situation, resulted in a four cycled 24 hour day. There were eight hours of diurnal conditions with the temperature at 86 F $\pm$ 2 and the relative humidity at 92%; a four hour crepuscular period for transition during which the temperature and humidity were altered to assume the nocturnal conditions of the tropical rain forest and the temperature at 72 F $\pm$ 2 and the relative humidity 92% to saturation. The nocturnal period was followed by another transition crepuscular interval and the cycle resumed. (Illustration 3).

Fungus population and activity within the tropical room was assured using banked beds of soil, decaying leaves, rotting cardboard, rotted equine and bovine feces and the walls were hung with untreated canvas duck. The atmosphere was circulated using fans and examined bi-monthly employing petri dishes of nutrient agar to define species population. The choice and adaptation of the room to missiles investigations was supported by data from previous experimental testing at Detroit and also from information forwarded to this installation from other places with similar equipment. (Illustration 4).

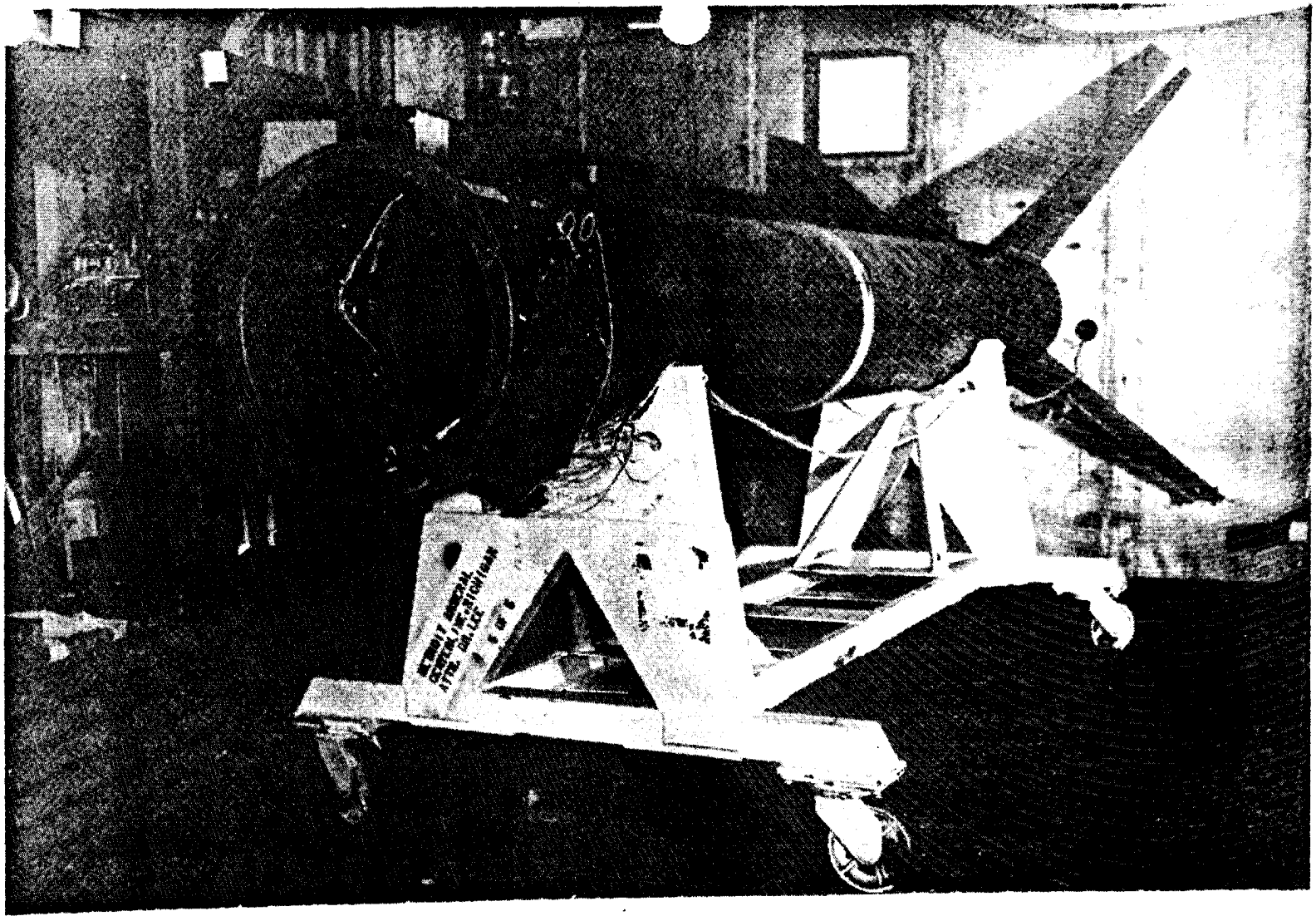
The use of the cycled atmosphere was evaluated by comparison and it was determined that with its use a greater number of species would be noted than using the room with constant temperatures and humidities.

In addition to the Tropical room, moist chamber cabinets were also used in the investigations because of the large amount of materials received. The constant situation substantiated the employment of cycled conditions with the wider spectrum of species and deterioration results than would be noted under the stable situations.



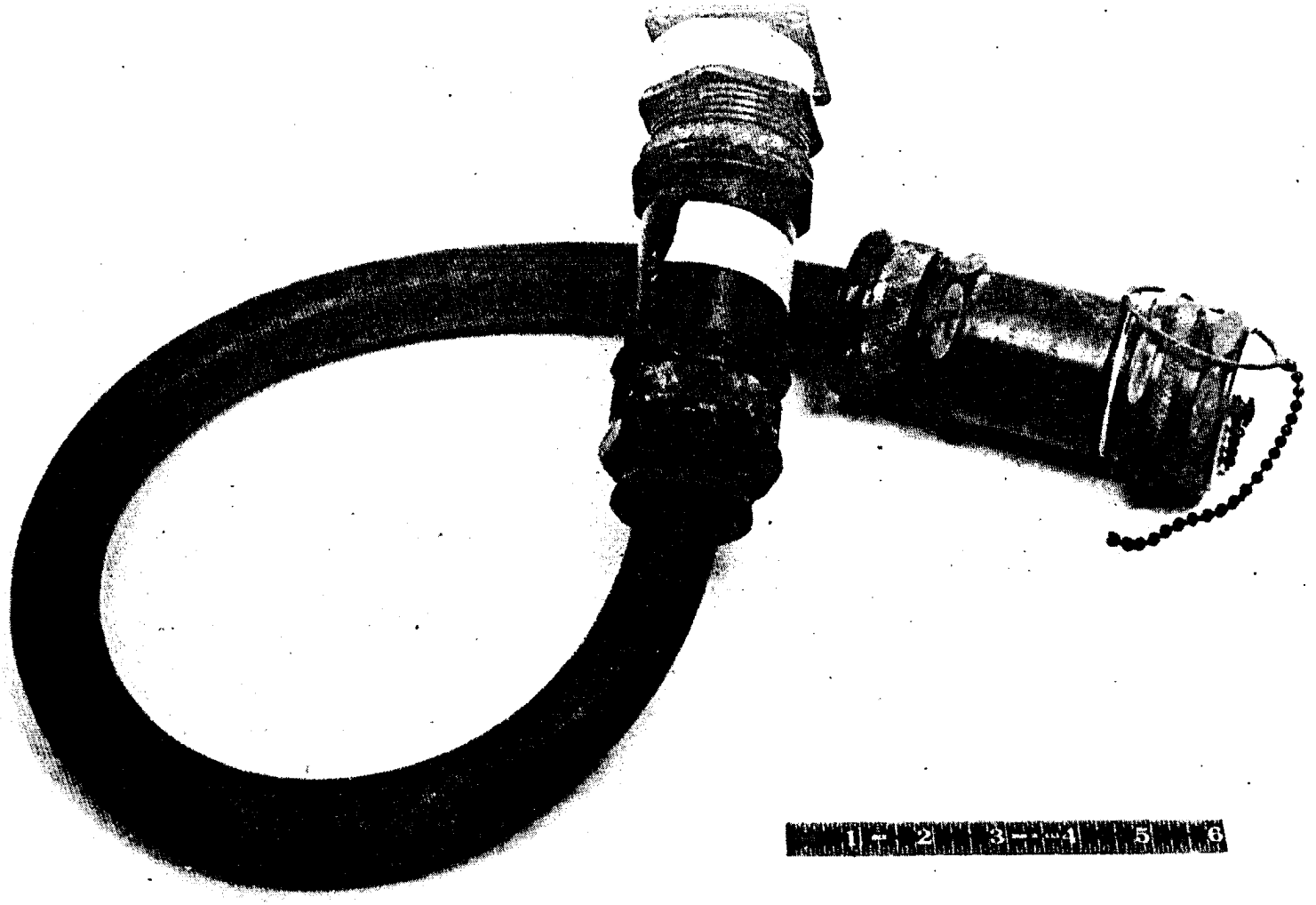
Graph of the twenty-four cycle of temperature and humidity conditions in the Tropical Room of the Detroit Arsenal - OTAC organization. An illustration of the diurnal, crepuscular and nocturnal conditions.

Illustration 4



The Honest John Rocket positioned for microbiological exposure in the Tropical Room at Detroit Arsenal - OTAC. Illustration includes racks holding components of the Redstone and also shows soil banked against the walls of the room.

Illustration 5



Multiconductor cable, FT-1, showing the importance of choosing funginert materials for inclusion in missiles. Cable photographed at the conclusion of 90 days of exposure within the Tropical Room at Detroit Arsenal - OTAC. There were no variations in pre-fungus and post-fungus exposure performance ratings.

The choice of testing situation indicated the third parameter to be followed in the microbiological testing of missiles; the importance of non-treatment of materials prior to testing. Past and present specifications often required a pre-treatment of materials be washing, placement into water baths with adjusted pH and temperatures, and the use of chemical cleaning, etc. All of these presented artificial barriers to securing accurate estimations of materials to fungus action. Would not it be more realistic and revealing of the actual resistance of materials to fungi if there was no pre-treatment of surfaces and compositions in any manner? Thus, for the eleven proposals of testing, no pre-treatment was employed, the materials being placed into the testing situation as received. (Illustration 5).

This parameter was not unique with missiles, but it was made official by inclusion into the missiles testing proposals and was chosen from data secured from testing on tank and tank-automotive vehicles. Employment of this parameter goes back to the idea of simulating in the laboratory as closely as possible, the conditions that would actually attain in storage or ready operation.

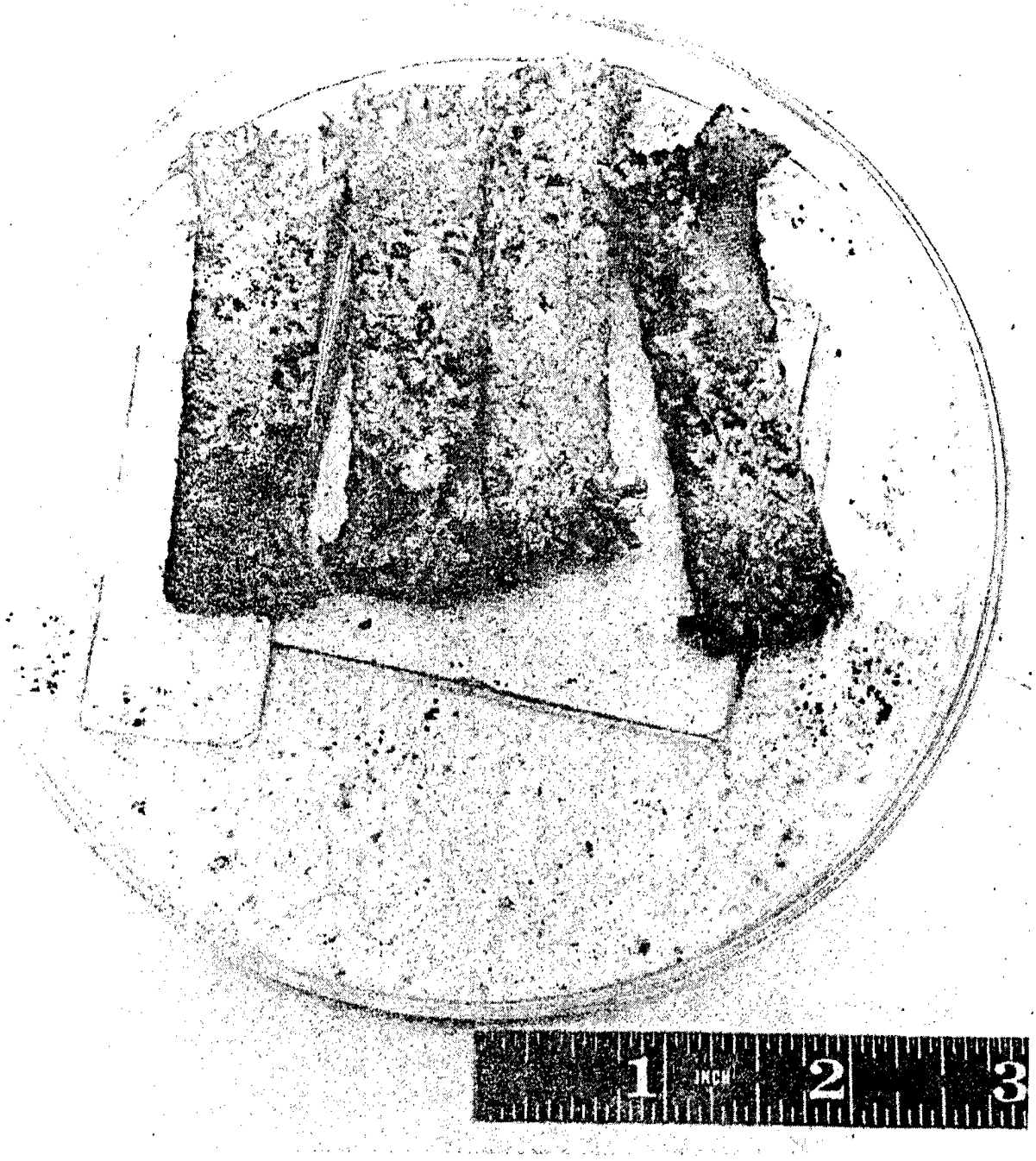
The fourth parameter developed for missiles testing was the choice of species of fungi. Again, in this matter, we relied on the data from past research and testing. However, since the majority of missiles items submitted for testing were new and unique, efforts had to be expended in securing information on composition of materials. Those which were cellulosic were inoculated with cellulolytic fungi; those proteinaceous with proteolytic species, etc. In the use of the various synthetics, the knowledge of the chemical composition was germane. In instances where it was impossible to define a material as to composition, a wide spectrum of fungus species was employed and the species mixture inoculated onto the material under test. At the conclusion of testing, observations were conducted to identify the fungi still evident and this information served as supporting evidence of actual utilization of the material. (Illustration 6).

The fifth parameter developed for the missiles research was the determination of the testing time. This is a crucial point and has been a concern of the Detroit organization ever since microbiology was established as a function.

Early specification tests prescribed a testing period of seven days. Later ones called for fourteen, with rarely twenty-one days. Since the first published specifications, testing time has gradually lengthened to ninety days. At Detroit, the shorter periods were viewed as unrealistic for producing data for estimating the effective resistance of materials' microbiological deterioration. Accordingly, over the years this laboratory has extended gradually the testing time of all components from thirty days to forty-five, to sixth, and currently ninety days. Only with the use of the longer period, it is felt, that we shall have a parameter to determine sufficiently the true behavior of materials to fungus attack. (Illustration 7).

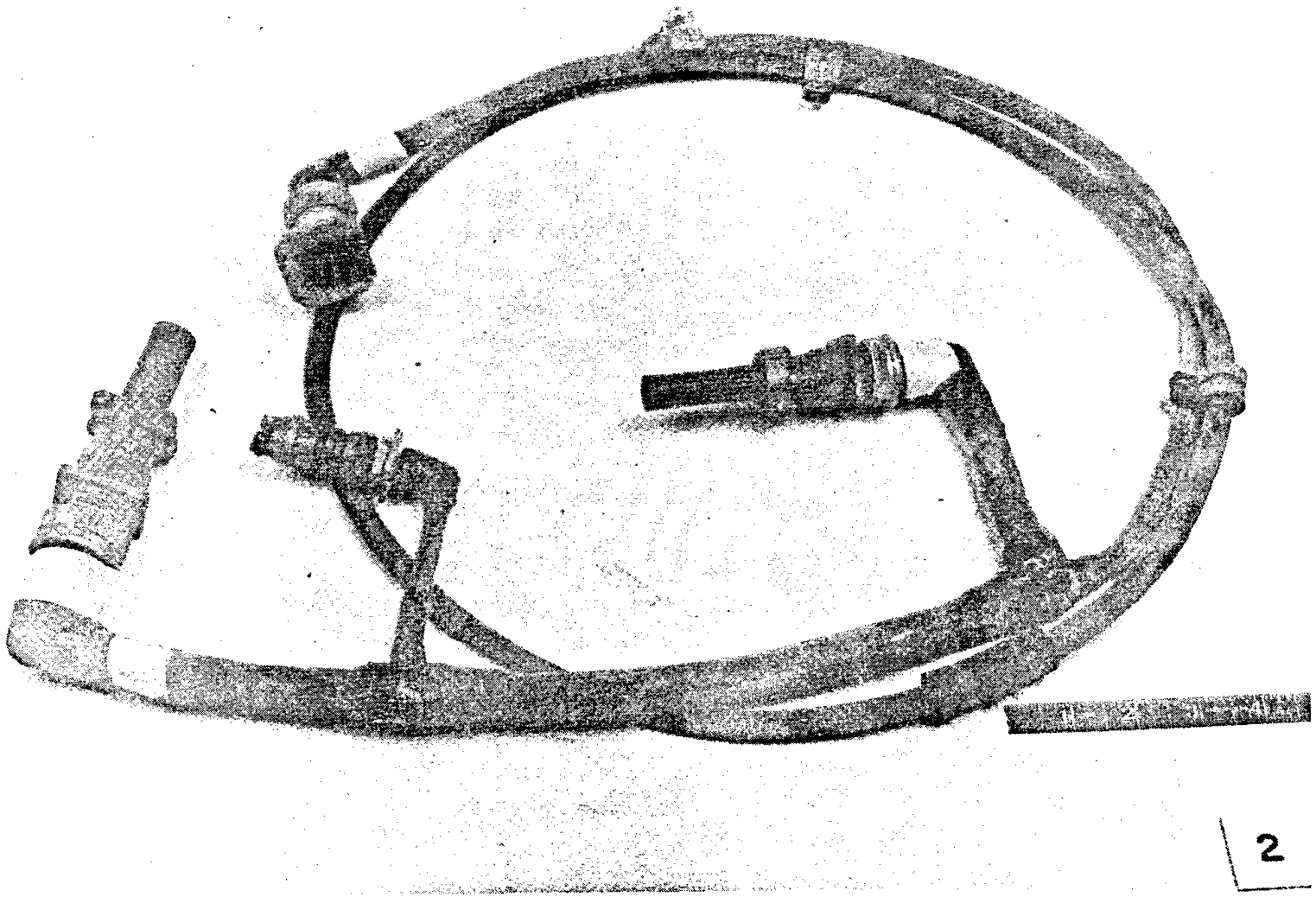
Support for the longer missile testing period relied on data secured on many dissimilar materials. In missiles work, it was noted that many of the items required a period for becoming conditioned to the atmosphere of the

Illustration 6



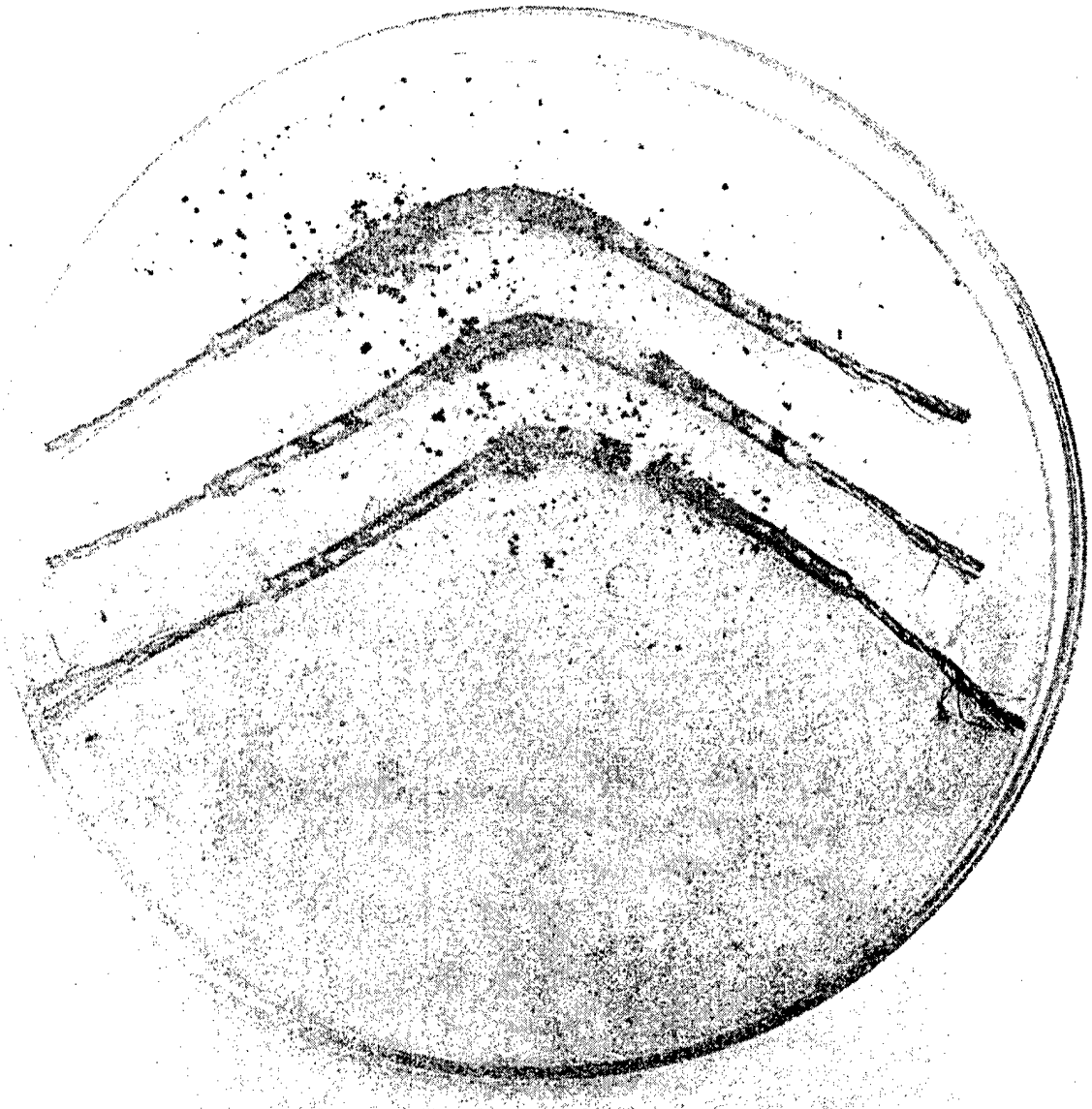
Asbestos samples adulterated with cotton. The cotton, being cellulose, supports a heavy growth of fungi while the funginert asbestos remains free of microbiological growth. Photograph taken at the conclusion of 90 days exposure in the Tropical Room at the Detroit Arsenal - OTAC.

Illustration 7

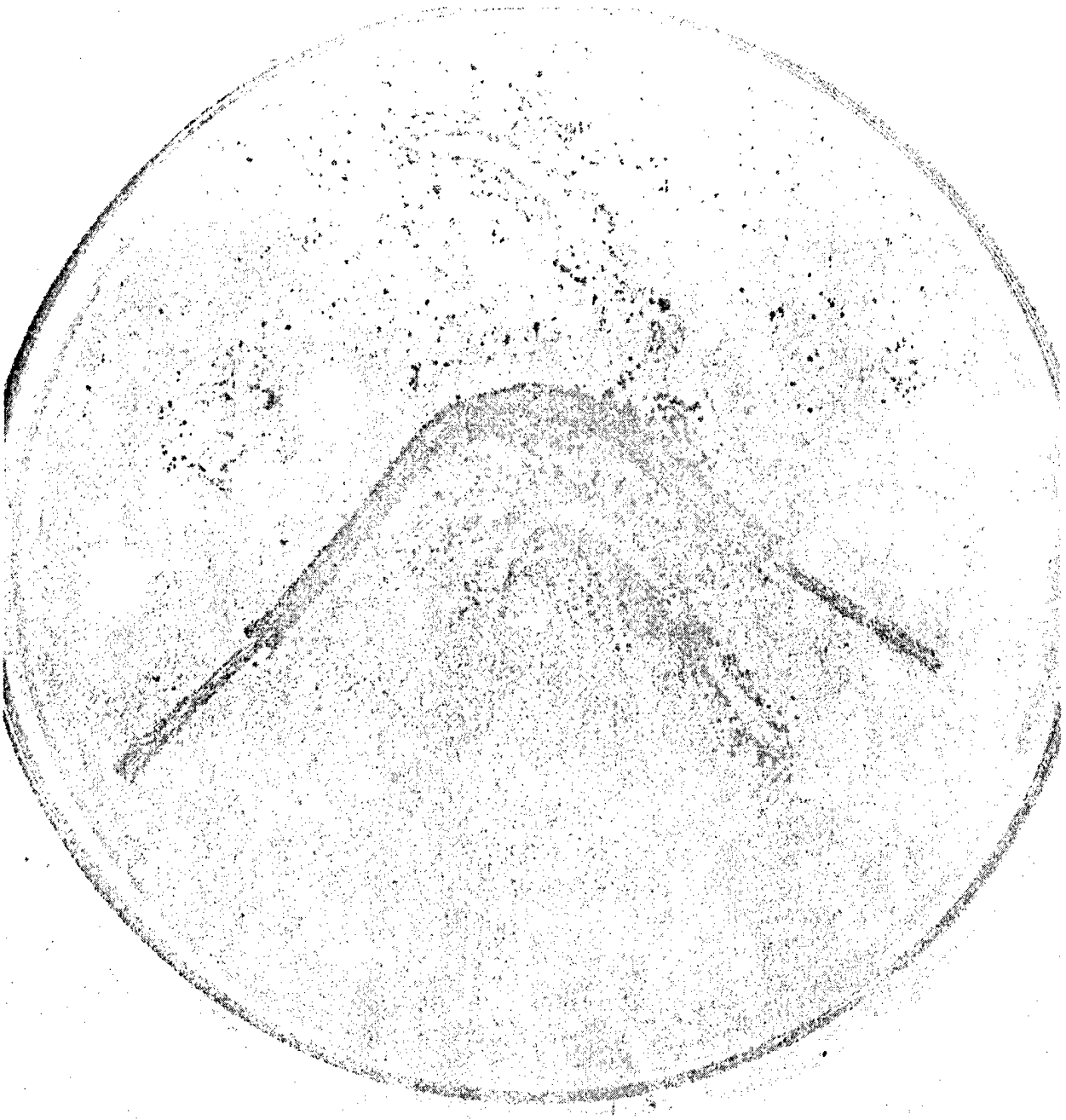


Cable harness following ninety days incubation with fungi in the Tropical Room of the Detroit Arsenal - OTAC. Cable insulation completely degraded and performance rating failed to meet the requirements for the component. Choice of inert materials is requisite.

Illustration 8

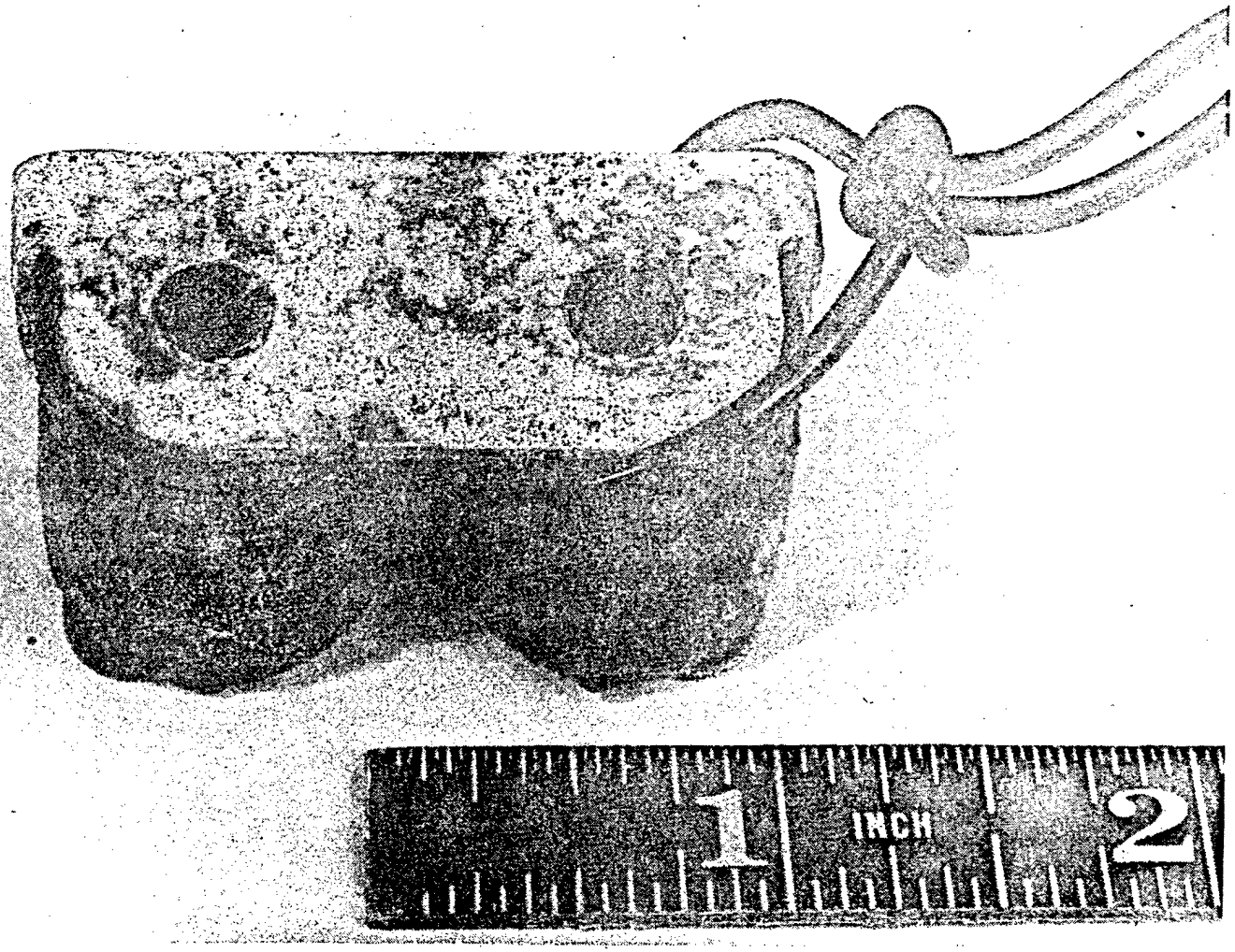


Cable leads following thirty days incubation in the Tropical Room of the Detroit Arsenal - OTAC organization. Both material resistance to fungus growth and performance ratings failed to meet minimum standards.



Cable lead at the end of ninety days incubation in the Tropical Room of the Detroit Arsenal - OTAC organization. Insulation completely degraded and performance failed to meet minimum requirements for the material.

Illustration 10



Summary of the weekly reports of a selected group of miscellaneous components used in the Redstone Missile, CCMD. Report covers a three months period and demonstrates the weekly changes in observations. Visual observations were also supported by performance ratings, where possible, at the thirty and sixty day periods.

testing chamber. Examination of missiles and components in storage for fungus development supported this contention. This adjustment period varied from thirty to sixty days and in that time there was often little development of fungi on missiles surfaces. However, once the materials were conditioned to the testing chamber, fungus growth proceeded rapidly and often apparently, instantaneously. This was particularly true of rubbers, plastics, and some of the miscellaneous components, assembled and disassembled components. The facts of the longer testing period accounted for the discrepancies also noted in our results as compared to other installations also performing tests on the same materials, but using shorter testing times. Fungi are living organisms and they all possess a threshold, above which, they cannot be stimulated to grow more rapidly. (Illustrations 8 and 9).

An adjunct of this longer testing period was the information found at the conclusion of the ninety day testing period. Materials removed from the testing chambers and placed onto tables out in the laboratory responded by developing different growth patterns, species of fungi developing, and loci on materials that were being utilized. This would not have been attained in the shorter testing periods.

An additional factor developed for demonstrating missiles resistance to fungus was the writing of the actual testing procedures. This was done whether the missiles were tested in toto or, disarticulated with parts disassembled. Testing followed basically the steps outlined in a specification developed in our laboratory and modified specifically for missiles and taking into consideration the parameters mentioned this morning. Consideration also had to be given whether missiles components were supplied as sealed or unsealed in an effort to eliminate a consideration for corrosion damage owing to moisture and which might have been primarily determined microbiological. (Illustration 10).

The use of the performance tests at this installation has been a parameter that has been pioneered at this place. All missiles materials, as received, were inspected for applicable, possible performance tests. Basically, these tests are demonstrations of the physical, chemical or mechanical properties and included data on strength, electrical conductance, depolymerization and visual evidence of changes such as complete or incomplete rotting, embrittlement, softening, bubbling, bleeding out of chemicals, crystallization of materials' surfaces, etc.

The use of the performance test was augmented and verified by the use of the periodic performance ratings secured from materials over the period of ninety days. These periodic tests were conducted within the tropical chamber so as to take advantage of the temperatures and humidities that would be found in the storage situations in the field. Further, these periodic tests verified the parameter of increased testing time. Often a material would fail within thirty days, while another would fail in sixty or at the terminal ninety days. The use of the periodicity in

testing allowed for savings in money and testing time by defining the exact time in which a material failed. (Illustrations 11 and 12).

This installation always requires a sufficient number of samples in order to allow for the periodic performance testing of items from preinoculation to final evaluation.

Fungus attack of missiles and missiles components is usually evident as surface growth on the various components. At our laboratory, the materials were separated prior to testing into coarse assemblages based on common characters. For example, we received:

natural and synthetic rubbers

electrical components, assembled

electrical components, unassembled

miscellaneous components containing plastics, finishes and textiles

coverings, insulations and gasketing

metallic units with organic-in-origin parts

single and multiconductor cables.

Fungus growth was noted on many of the preceding. However, it was not employed as a definitive parameter without the supporting data from other parameters previously mentioned. Visual evidence is deceptive and decision is required whether the growth is adventitious or deleterious. Using the performance test, the latter is easily accomplished and data based on changes in physical properties such as losses in tensile strength, powdering of surfaces, scuff resistance alterations, loss or increase in adhesion; chemical tests with alteration in composition or electrical measurements of changes in current carrying capacity. All of the preceding, of course, require a comparison with pre-fungus test data in order to have a comparison with the post test ratings.

All of the information presented this morning has been considered in forming conclusions on the importance of fungi as deteriorating agents on missiles and missiles components. Further, the results of our investigations indicated that control of microbiological deterioration is necessary in order to eliminate fungi as possible causative factors of malfunction from the manufacture to ready storage and ultimate operation.

Illustration 11

Table I – SUMMARY OF WEEKLY REPORTS (Continued)

		Asbestos Tape	Cotton-Vinyl Tape	Wool Felt	Wool Felt	Asbestos Binders	Insulation Sleeving	Silicone Compound	Pressure-Sensitive Tape	Silicone Compound	Cotton Tape	Vinyl Tape	Foam Rubber	Vinyl Coating
Sample Number		14	15	16	17	18	19	20	21	22	23	24	25	26
Report Period	Exposure (Days)													
14-18 Jul	7	*	S	*	M	*	M	*	*	*	*	*	M	XX
21-25 Jul	14	*	SM	*	SM	M	SV	*	*	*	S	S	MS	XX
28 Jul 1 Aug	21	*	*	S	SV	M	E	*	*	*	S	S	SV	*
4-8 Aug	28	*	*	S	SV	M	E	*	*	*	S	S	E	S
8-14 Aug	35	*	SM	S	E	M	E	*	*	*	S	S	E	M
14-21 Aug	42	*	M	M	H	M	E	*	S	*	S	S	E	SV
21-28 Aug	49	*	H	M	H	M	E	*	S	*	S	S	E	E
29 Aug 4 Sept	56	*	H	H	H	M	E	*	SV	*	S	MS	E	E
5-11 Sept	63	*	H	E	H	M	E	*	SV	*	S	E	E	E
12-19 Sept	70	*	H	E	H	M	E	*	SV	*	S	E	E	E
19-26 Sept	77	*	H	E	H	M	E	*	SV	*	S	E	E	E
27 Sept 3 Oct	84	*	M	E	H	*	E	*	SV	*	SM	E	E	E
4-10 Oct	91	*	M	E	H	*	E	*	SV	*	SM	E	E	E

XX-Sample not available for starting date; however, growth was extensive at the end of the test.

* No fungus growth
S Slight fungus growth
SM Slight-to-moderate growth
M Moderate growth
MS Moderate-to-severe growth
SV Severe growth
E Extensive growth

Magnetic counter, solenoid coils, from improperly sealed component. Photograph taken at the conclusion of ninety days and indicates the importance of correct sealing of components in the elimination of fungus attack of material

Illustration 12

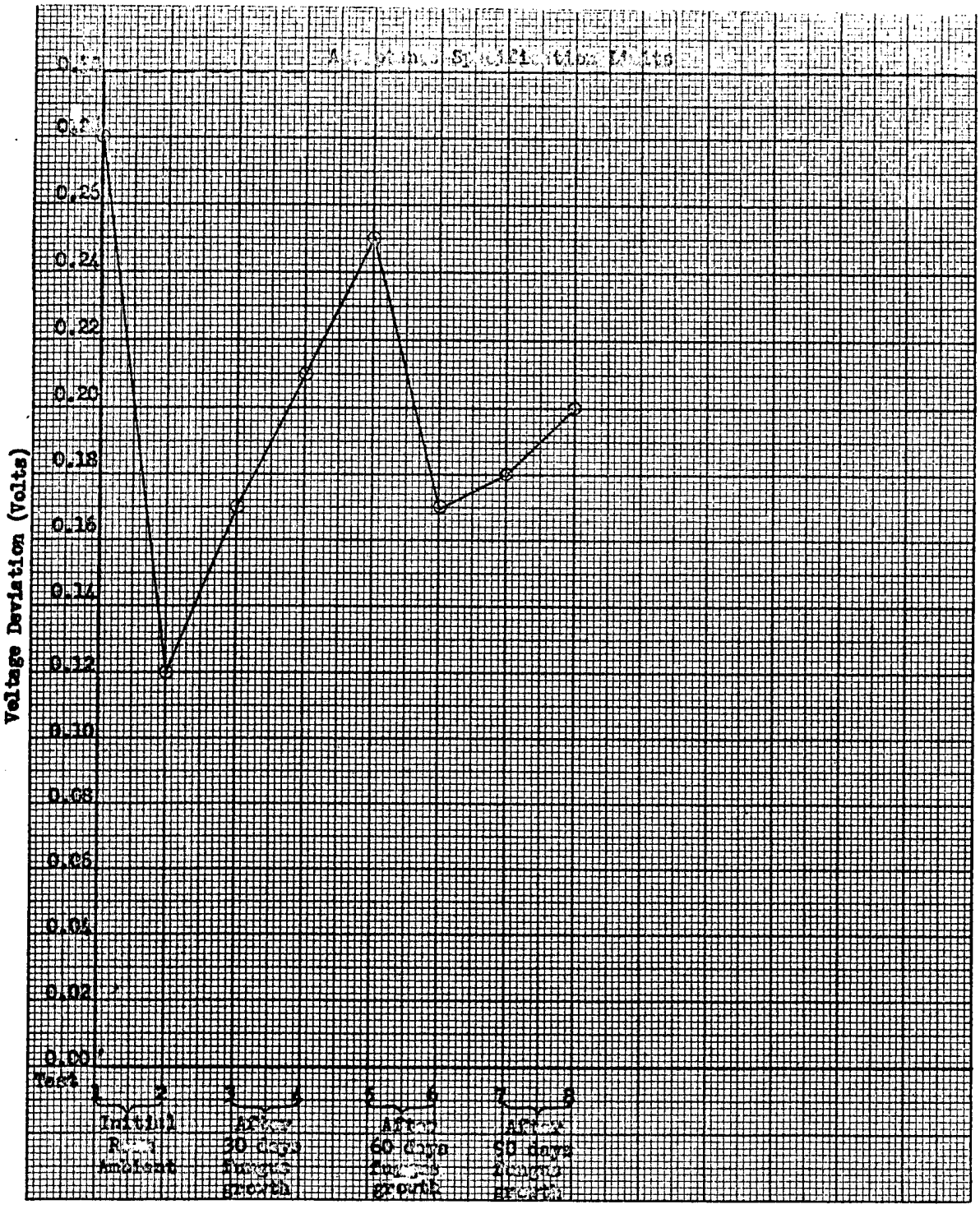


Figure 7 - Voltage deviation for constant-input, constant-load test (two-hour period)

The importance of performance testing before, during and at the completion of microbiological exposure. Voltage deviation

DESIGN OF ENVIRONMENTAL EXPERIMENTS AND RELIABILITY PREDICTION

A. Bulfinch
Feltman Research and Engineering Laboratories
Picatinny Arsenal

. . . ——————— . . .

PREFACE. This paper demonstrates how accepted statistical techniques stand to reduce the cost of testing missiles and missile components. These techniques are not treated in full; however, examples of their use and references are given.

The subject matter covered is oriented towards practising reliability engineers. It is hoped that some of their troublesome problems have been clarified.

ABSTRACT. Describes the need for, and use of factorial designs in surveying the separate effects of a large number of environmental treatments with maximum reliability and a minimum number of test specimens. Shows how information from the factorial experiment can be used to define reliability, and how this information can be used in tests of increased severity to predict "reliability-in-use."

SUMMARY. Reliability is defined.

The need for testing to failure is emphasized by comparing construction engineering problems with missile reliability problems. Methods are given to convert "safety margins" to measures of probability which can be used to predict "reliability-in-use."

The advantages of using factorial designs to survey the separate effects of a large number of environmental conditions are described. Information obtained from factorial experiments is used to formulate the definition of reliability in terms of the severest environments. These environments are used in tests of increased severity. Tests of increased severity are used to establish the relationship between use conditions and test conditions in order to predict "reliability-in-use."

Detailed examples are given of methods of predicting high "reliabilities-in-use" with small sample sizes.

CONCLUSIONS.

1. Any particular component can have many reliability values simultaneously. There is one reliability value associated with each possible combination of environmental condition and measurable functioning characteristic.

2. The results of factorially designed environmental experiments should be used in defining component reliability.

3. Tests of increased severity should be used in combination with factorial designs to predict "reliability-in-use" from test results.

4. Tests of increased severity can demonstrate high "reliabilities-in-use" with small sample sizes.

RECOMMENDATIONS.

1. Factorial designs should be used in combination with tests of increased severity to predict "reliability-in-use."

2. The test condition used should be experimentally determined.

3. Military standards should be revised to permit the experimental determination of the test conditions to be used.

4. The terms on which reliability is defined should be experimentally determined.

. . . ——————— . . .

I INTRODUCTION

A General

The statistical aspects of reliability are not new. All of the necessary concepts are adequately treated in modern statistical literature. The lack of information about measurable characteristics of the missile system and the environment it experiences in use, as well as the high cost of test specimens, have created the current problems.

Urgently needed are highly efficient, experimental techniques that can be uniformly applied by various segments of the same organization and by different organizations. Highly efficient techniques are required because of the need to demonstrate very high reliabilities with very small sample sizes. High reliabilities are required, of course, to assure succesful functioning of complex systems composed of many components. Only small samples can be used because of the high cost and/or scarcity of test speciment. Uniform, standardized procedures are required for the collection of comparable data.

B Purpose

This paper purposes to do the following:

1. Define component reliability.

2. Describe how factorial designs can be used to survey the effects of several environmental conditions (with minimum sample size) preparatory to defining reliability in terms of these conditions.

3. Describe how "reliability-in-use" can be predicted from results of laboratory tests of increased severity.

4. Describe how tests of increased severity can use information from factorial experiments.

C Scope

The accepted statistical definition of reliability follows: Reliability is "the probability of successful functioning in use." This is a general definition that is applicable to any operating system. To define component reliability, the general definition must be modified to include:

a. Environmental conditions under which successful functioning took place.

b. The component characteristics that functioned successfully.

This means that every component can have as many reliabilities as the total number of possible combinations of environmental conditions and component characteristics. To have the weapon system reliability meaningful, the component reliabilities should be defined in terms of the most severe conditions so that the stated component reliability will be the minimum.

D Background

The techniques described, the most efficient known, are designed to maximize the amount of information obtainable from a given sample size. In addition, these techniques are definitive enough to serve as standard procedures throughout the same or different organizations over an extended period of time.

The uniform applicability of these techniques is as important as the efficiency. A large part of the value of experimentally determined reliability data is their scope of applicability. That is, reliability data collected by means of standardized procedures are cumulative in the mathematical sense. Hence, the precision with which reliability values are known can be improved with time as additional data are collected. This makes it possible to collect a reference file of reliability data on a variety of standard components.

II STATEMENT OF PROBLEM

In the usual case, the development engineer has one or more items to test under many different environmental conditions. The items to be tested may have two or more properties that must be evaluated.

Objectives of the usual reliability experiments for newly developed items follow:

1. Determine how well the engineer has succeeded in developing a reliable item.

2. Obtain an unbiased estimate of the "strength" (i.e., ability to withstand stress) of the item with minimum cost.

3. Determine the separate and combined effects of the environmental treatments on the reliability of each property measured.

4. Determine the effect of the length of time under the separate and combined effects of the environmental treatment on the reliability of each property measured.

5. Predict the "reliability-in-use" from the test results.

Objective 4 requires life-testing techniques which have been treated extensively elsewhere. This paper describes other techniques for increasing test severity.

The multiple properties of an item can be measured simultaneously. This poses no particular problem. The real problem confronting the engineer derives from his having different components that must be treated with a large number of different environments on a very limited budget.

Since most components are unique, they must be treated as separate problems. There is no known way to combine different kinds of components into a single integrated experiment. However, a single integrated experiment can be designed for several kinds of environments. As a result, the problem is one of designing the most efficient experiment for a single type of item and repeating the process for each type.

It is assumed that in every missile component there exists a true but unknown "strength," created by the particular design developed and used by the engineer in building the component. It is further assumed that the true "strength" is a constant and not a random variable for any particular design over short periods of time.

The present practice of designing components to pass the current military standards without failure does not attain the intended objectives for the following reasons:

1. An unbiased estimate of the true strength of the item cannot be obtained unless some items fail.

2. Cost prohibits testing all items at the same level of severity.

3. Reliabilities are demonstrated only in proportion to the number of items tested.

Testing without failure all items at the same level of severity can lead to completely erroneous conclusions. For example, in comparing two designs, there may be available more test specimens of the poorer design. Tests by military standards result in an equal number of failures for the two designs. Under these conditions, experimental evidence favors the poorer design.

Instead of subjecting all test specimens to the same test conditions regardless of their intended use, the level of severity of the test should be progressively increased until a failure is obtained. This procedure will lead directly to an unbiased estimate of the true strength with a minimum number of test specimens and establish the correct level of severity for each type of component at which the failure rate should be determined. This procedure will correctly differentiate between different designs.

Experience in the construction engineering field has shown that assurance that an item will not fail in use requires a large "safety factor" to be built into the item. To determine the "safety margin," the load applied must be increased until the test item breaks, or fails in some other manner. This is, in effect, a test of increased severity that leads directly to an estimate of the true breaking load the engineer is seeking to determine. The average, and standard deviation of only a few (3-6) such results are all that are required, because each value so obtained is an estimate of the true value. The difference between the "observed average breaking load" and the "load expected in use" divided by the "load expected in use" is called the "safety margin" or margin of safety. The larger this value is, the "safer" the engineer feels in predicting that the item will do the intended job without failing.

The construction engineer could have elected to load each test specimen with only the load expected in use, but because he has designed the item to withstand the load expected in use, this procedure tells him nothing about the true breaking load. All he learns is what he already knows--that it will not fail! Now if he wants to "feel safe" in predicting invariably successful functioning of the item, he must test many items. Asked to conduct his test in this manner, the engineer would rebel because he knows, as we do, that--it is far too costly!

In missile component testing, we should simulate the procedure used by the construction engineer-"load the item until it breaks" and then calculate the "safety margin." To do this, we must shift our attention from finding the number of items to be tested without failure at a single level of severity to finding the level of severity that will cause failure, and then finding the failure rate at that level. That is, to find the reliability (the probability of success) we must first find the failure rate (the probability of failure). To do this with small sample sizes, we must use a test of increased severity to find the level that will cause failure.

Robert Lusser, formerly of Redstone Arsenal, has advocated this approach of "testing to failure" for some time (Ref. 1). However, he neither showed how to "load" missile components until they fail, nor

predicted the "reliability-in-use" by means of the laws of probability. He was satisfied with using "margins of safety" (Ref. 2).

III APPLICATION OF PROBABILITY LAWS

An item will not fail until the applied stress exceeds the item's "strength." If the "strength" is much greater than the stress expected to be experienced in use, the chance (probability) of failure in use is very small, and the chance of success (reliability) is very high. It is in this sense that "high reliability" is defined. That is, high reliability means high probability of successful functioning under actual use conditions; it does not mean high reliability under the test conditions.

To translate the reliability demonstrated under test conditions to a "reliability-in-use" value, the relation between the "use and "test" conditions must be known. Experience has shown that this relationship can be adequately represented by frequency distributions. This places the relationship on a probabilistic basis, and makes possible the use of the laws of probability.

When the average of the conditions in use is known, the level of severity required at which items must be tested to demonstrate any given reliability can be calculated in advance. As a result, reliabilities can be correctly predicted with small sample sizes without testing to failure. Alternatively, when the ultimate strength of the item is desired, the first failure method described below can be used. Both of these procedures predict "reliabilities-in-use" with small sample sizes through the use of the multiplication law which states that the probability of simultaneous occurrence of two independent events equals the product of the probabilities of separate occurrences of the events. Examples of both of these procedures are given below.

In reliability testing, the two simultaneous events referred to are test specimen failure and test condition causing that failure. Both the "failure rate" and the "chance" of the test condition's occurrence in "use" can be considered probabilities. By the above law, the predicted failure rate in use will be the product of the failure rate obtained in testing and of the chance that the test conditions could occur in use. The predicted "reliability-in-use" will then be equal to one minus this product.

When nothing is known about the environmental conditions expected in use, or when these conditions may vary in an unpredictable way, no prediction can be made about a unique "reliability-in-use." However, these methods of testing to failure can still be used to advantage. Knowing how much punishment components can take before failing reduces the number of unknowns, can be valuable in deciding how, and under what conditions, a particular missile can be used. This information can be useful in choosing between missiles for particular purposes. Moreover, where nothing is known about the conditions in use, the "most severe use condition" the item can be subjected to for any specified reliability can be calculated. An example is given below.

The methods described below show how missile components can be loaded (tested) to failure through the use of tests of increased severity. These methods also show how safety margins can be used to predict "reliability-in-use."

IV LABORATORY TEST METHODS

It is assumed in these methods that the test item can fail in but one way; that is, the binomial distribution is applicable.

A Factorial Designs

Plans should be made to conduct the laboratory experiments in two stages. First, survey the separate effects of the several environmental conditions of interest in one integrated experiment. The two-to-the-n^{th} factorial designs or their optimized modifications are the most efficient for this purpose. These designs can be used to select the treatments causing the highest failure rates. These treatments can then be used to define the reliabilities of the test item. If the reliabilities determined in terms of these treatments are acceptable, the reliabilities of the test item in terms of any of the other treatments will also be acceptable. This procedure will reduce the magnitude and complexity of the experiments conducted to determine and predict reliability. More importantly, component reliabilities obtained in this manner will furnish a more realistic basis for calculating systems reliability.

See references 5 and 6 for available designs. These designs are the most efficient known. Experiments based on these designs may be conducted without changing the treatment procedure except to arrange for the test specimens to receive the number and kind of treatments required by the particular design used. However, the best differentiation among treatments is obtained when the level of severity used will cause 50 percent of the test specimens to fail.

For the purpose of this application, the two levels of each treatment can be the presence and absence of the treatment. Alternatively, any two levels of the treatments can be used.

The number of test specimens required in the optimized designs is one more than the total number of treatments used (Ref. 5). The more versatile fractional factorial designs (Ref. 6) require at least 16 items for experiments containing from five through eight treatments, and at least 32 items for nine through 13 treatments. With twice these numbers of items, the latter type designs can also measure interactions--how the effect of any one environment depends upon the others. Interactions among treatments cannot be measured by any design except the factorial.

Factorial designs permit a type of statistical analysis that distinguishes between variations due to chance and variations having assignable causes, and produces more information from a given number of items than any other known procedure. These designs actually

increase the effective sample size by making it possible to use each observation (or measurement) for more than one purpose. In fact, each treatment effect is determined as though the entire experiment is conducted to determine that particular treatment effect alone. As a result, the reliability with which each treatment effect is determined can be based on the total number of items used in the epxeriment. The three-treatment-design example described below demonstrates this point.

Further advantages in using factorial designs in environmental testing experiments follow:

a. No control groups are required.

b. Each treatment effect can be determined independently of all the others. That is, unambiguous conclusions can be drawn about each treatment effect.

c. Complex experiments involving a large number of treatments can be easily handled with the factorial procedures.

d. This is the only experimental design in which the relationship among the treatments can be measured. That is, the factorial design can determine whether the effect of one environmental treatment depends upon any of the others. These effects are called interactions.

e. The probability of being right or wrong can be controlled.

f. When the number of treatments used becomes large (three or more), only a fraction (1/2, 1/4, 1/8, etc.) of the total number of combinations in the factorial need be used.

When multiple replications cannot be used and only attribute (go, no-go) data are available, these designs can still be used to take advantage of their efficiency. However, in cases of this kind, the usual analysis of variance cannot be made. Instead, the usual summations are made to obtain and compare two binomial proportions (by the Fisher exact method) to determine the effect of each treatment. See Example No. 1 below.

Results of factorial designs are used as a guide in determining how to define reliability prior to conducting the test of increased severity. That is, the factorial experiment surveys all of the environmental treatments of interest (with a minimum number of test specimens) to determine the difference, if any, among the environmental effects. A decision is then made whether to redesign the item. If the item is considered acceptable at this time, reliability is defined in terms of the environmental treatment or treatments found to be most severe. If no differences are found among the effects, reliability can be defined in terms of a combination of several of the treatments considered most important from an engineering point of view. If reliability is defined in terms of the most severe treatments, the reliability values obtained will be smaller than those obtained with the other treatments. This is a necessary condition if the system's reliability derived from the component's reliabilities is to have meaning.

Tests of Increased Severity

Results obtained from the factorial experiments can be used to determine which of the environmental treatments will be used in the following procedures to predict "reliability-in-use."

B First-Failure Method (Single Factor)

Increase the level of severity after each test result is obtained until the test item fails. If the test destroys the item, increase the level of severity used with each succeeding item tested until a failure is obtained.

The level of severity can be increased in a variety of ways, such as the following:

1. Using more extreme treatments (e.g., higher temperatures or higher G-values).

2. Using two or more treatments on each test specimen.

3. Repeating the same treatment or set of treatments on the same item.

NOTE: Increasing the length of time an item is subjected to a particular treatment is not used here as a means of increasing the degree of severity.

By starting at, or near the level of severity expected in use, a failure should be obtained within five or six trials (or items). After the level of severity has been found that will cause failure, three or more items should be tested to estimate the failure rate at this level.

To determine the predicted "reliability-in-use," find the probability of occurrence in use of the test condition (at which the failure rate was measured) from a table of individual terms of the Poisson distribution (such as Table 39 of reference 9), where "m" is the expected use condition used. This in effect determines the probability associated with the "safety margin." The product of this probability value and the failure rate found under the test condition is the predicted failure rate in use. The predicted "reliability-in-use" is equal to one minus this product.

V EXAMPLES

A Example No. 1

This example demonstrates how factorial designs can be used in combination with tests of increased severity. A simple three-treatment-experiment example is given below. The treatments used in this example are identified and defined as follows:

Identification	Treatment
A	Trans. Vib.
B	Flight Shock
C	Waterproofness

For purposes of the factorial design, each treatment is considered to have two levels:

1. Lower level or absence of the treatment (designated by subscript 1).

2. Higher level or presence of the treatment (designated by subscript 2).

The total number of possible combinations of three treatments, each at two levels, is two cubed or 8. These 8 combinations can be written in the following pattern:

	A_1		A_2	
	B_1	B_2	B_1	B_2
C_1	(1)	b	a	a + b
C_2	c	b + c	a + c	a + b + c

A minimum of 8 items would be required for this plan, each receiving different treatment combinations as follows:

Item Number	Treatment Combinations
1	None (1)
2	B only
3	A only
4	A + B
5	C only
6	B + C
7	A + C
8	A + B + C

By using the same letters (A, B, and C) and symbol (1) to represent the results obtained from testing the eight items, it can be shown symbolically that the treatment effects can be independently determined, using the total number of items in the entire experiment for each treatment as follows:

EFFECT OF TREATMENT A

$$A + (A + b) + (A + c) + (A + b + c) - [(1) + b + c + (b + c)] = 4A$$

EFFECT OF TREATMENT B

$$B + (B + c) + (a + B) + (a + B + c) - [(1) + c + a + (a + c)] = 4B$$

EFFECT OF TREATMENT C

$$C + (b + C) + (a + C) + (a + b + C) - [(1) + b + a + (a + b)] = 4C$$

One-fourth of these differences equals the average effect of the respective treatments. From the above equations, it can be seen that the results obtained from the eight items have been used three times--once for each treatment. This produces an effective sample size equal to 3 X 8, or 24 items; yet, each treatment has been determined independently of the others.

The above three-factor factorial can be used as an example of a fractional factorial design as follows:

	A_1		A_2	
	B_1	B_2	B_1	B_2
C_1	-	b	a	-
C_2	c	-	-	a + b + c

A minimum of four items is required in this design. As before, the separate effects can be determined by a process of summation and subtraction as follows:

EFFECT OF TREATMENT A

$$A + (A + b + c) - (b + c) = 2A$$

EFFECT OF TREATMENT B

$$B + (a + B + c) - (a + c) = 2B$$

EFFECT OF TREATMENT C

$$C + (a + b + C) - (a + b) = 2C$$

One-half of these differences equals the average effect of the respective treatments.

When there is only one item available for each treatment combination, and only success and failure data are available, the usual analysis of variance cannot be used. However, the above differences, which will be binomial proportions in this case, can be compared by the Fisher exact method for 2 x 2 tables (Ref. 7) to determine the treatment effects. A very convenient set of tables for this purpose can be found in Ref. 8, which contains tables of minimum contrasts based on Fisher's exact method.

When it can be determined, from the results of the factorial esperiment, which environmental conditions will be used to define reliability, the level (or severity) of the condition required to demonstrate a given "reliability-in-use," with a small sample, can be calculated in advance of testing, on the assumption that no failures will be obtained; if the average condition in use is known.

The test conditions required can be calculated as follows:

$$R = 1 - P_t \text{ (UCL)}$$

where:

R = the specified "reliability-in-use."

P_t = the probability of test conditions' occurring in use.

UCL = the upper confidence limit (associated with the specified confidence level) of the failure rate expected under the test conditions to be calculated below.

When R and UCL are known, P_t can be calculated from the above formula. Given P_t (the probability) and the average use condition (m), the required test condition (i) can be found from Table 39 of reference 9, or from the following formula:

$$P_t = m^i e^{-m}/i\ !$$

where (e) is the base of natural logarithms.

1 Sample Calculations

Using the same three-factor-experiment example as above gives the following typical set of results, when one is entered as a "failure" and zero is entered as a "success." It is assumed that a knowledge of the item being tested has led to the decision that transportation vibration, flight shock, and waterproofness, in that order, are the three environmental conditions most likely to affect the important functioning characteristic of this item; this characteristic is contact resistance.

The treatment procedure and work-sheet (to record results) for this experiment would be the following two-entry table. An "X" in the item column means that the item receives the corresponding treatment, while a blank means that the item does not receive the treatment.

Treatment Procedure

Order of Treatment	Item No.							
	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32
Trans. Vib. (A)			X	X			X	X
Flight Shock (B)		X		X		X		X
Waterproofness (C)					X	X	X	X
Results: Replication 1	0	0	1	1	1	1	0	1
2	0	0	1	0	1	1	0	0
3	0	1	1	1	1	1	0	1
4	0	1	0	1	0	1	0	0
Totals	0	2	3	3	3	4	0	2

The results of one complete replication should be obtained under a single set of controlled conditions (e.g., in the same day, same operators, same instruments, etc.), before going to the next replication. This will make it possible mathematically to subtract out of the results the effect of changing conditions.

By placing these results in the usual factorial matrix, the following table would be obtained:

	A_1		A_2	
	B_1	B_2	B_1	B_2
C_1	0	0	1	1
	0	0	1	0
	0	1	1	1
	0	1	0	1
	0	2	3	3
C_2	1	1	0	1
	1	1	0	0
	1	1	0	1
	0	1	0	0
	3	4	0	2

In preparation for analyzing these results, the usual summing process would give the following series of two-factor tables:

Summing over A

	B_1	B_2	Row Totals
C_1	3	5	8
C_2	3	6	9
Column Totals	6	11	17

Summing over B

	A_1	A_2	Row Totals
C_1	2	6	8
C_2	7	2	9
Column Totals	9	8	17

Summing over C

	A_1	A_2	Row Totals
B_1	3	3	6
B_2	6	5	11
Column Totals	9	8	17

Each one of the marginal totals is the sum of 16 observations. The results can now be analyzed and interpreted as follows:

Source	Effects	Test of Significance [a]
Main Effects		
Trans. Vib. (A)	9/16 vs 8/16	non-significant
Flight Shock (B)	9/16 vs 11/16	non-significant
Waterproofness (C)	8/16 vs 9/16	non-significant
Replication		
1	5/8	non-significant
2	3/8	non-significant
3	6/8	non-significant
4	3/8	non-significant
Interactions		
A x B	8/16 vs 9/16	non-significant
A x C	4/16 vs 13/16	significant
B x C	8/16 vs 9/16	non-significant
A x B x C	7/16 vs 10/16	non-significant

[a] By the Fisher exact method for the 95% (two-sided) confidence level

2 Interpretation (when the above order is used)

(a) None of the effects is significant except the AC interaction. This means that items which have received transportation vibration treatment are significantly less waterproof than those not receiving transportation vibration.

(b) None of the treatments taken alone is significant, although the flight shock effect approaches significance. This result suggests the need for additional flight-shock tests if this treatment is considered important from an engineering point of view.

(c) The fact that the three-factor (ABC) interaction is not significant shows the following:

(1) Waterproofness does not change the effect of transportation vibration on flight shock (AB interaction).

(2) Flight shock does not change the effect of transportation vibration on waterproofness (AC interaction).

(3) Transportation vibration does not change the effect of flight shock on waterproofness (BC interaction).

(d) The fact that replication is not significant means that conditions were under a state of control throughout the experiment.

These results show clearly that the effect of transportation vibration on waterproofness is the most severe combination. It would appear from the results that a decision to improve the waterproofness characteristics is required. After this has been done, reliability must be defined. The results of this experiment show that reliability should be defined in terms of contact resistance (the functioning characteristic of interest) under the following environmental conditions:

(1) Transportation vibration followed by waterproofness (since these two conditions interact).

(2) Flight shock (since this treatment effect approaches significance).

If the reliability of the contact resistance under these conditions is acceptable, the reliability of the contact resistance under the other conditions will also be acceptable.

If the average transportation vibration condition in use is assumed to be 5 G's and the required reliability is 0.995, the test condition required to demonstrate this reliability with a sample of 5 test specimens can be calculated as follows:

when: $R = 1 - P_t$ (UCL)

then: $P_t = \frac{1 - R}{UCL}$

when: $R = .995$

$UCL = .52$ (the upper confidence limit at the two-sided 95% confidence level for testing five items and obtaining no failures)

then: $P_t = \frac{1 - .995}{.52} = .0096$

From Table 39 of reference 9 the test condition (i) associated with a use condition (m) of 5 G's and a probability (P_t) of .0096 is found to be equal to 10.9 G's. This is the level of transportation vibration required, followed by the waterproofness test to demonstrate a contact resistance reliability of 0.995 if no failures are obtained. If failures are obtained, the test conditions required to demonstrate a reliability of 0.995 with a sample of five test specimens are as follows:

Observed No. of Failures in Sample of 5 Test Specimens	% Failure	Test Condition Required, in G's
0	0	10.9
1	20	11.3
2	40	11.5
4	80	11.7

NOTE: It is evident from the above sample calculations that, for small-sample sizes, the difference in test conditions between zero and anything less than 100% failures is insignificant. This means that the required test condition can be conservatively estimates by expecting a high failure rate.

B Example No. 2

When the average of the conditions in use is known, very high values for the "reliability-in-use" can be correctly predicted with very small sample sizes, if the level of severity is increased until a failure is obtained:

Given:

Average use conditions (m) = 5 G's

Found (Using First - Failure Method):

Number of items used to find test condition to cause first failure	5
Test condition found (i)	18 G's
Number tested at 18 G's	5
Number of failures at 18 G's	2

The probability (P_t) of the test condition's occurring in use, from Table 39 of reference 9, when m = 5 and i = 18, is found to be $P_t = 0.000004$. The upper confidence limit (UCL) of the observed failure rate (2/5) for the two-sided 95% confidence level, from Table V of reference 8, is found to be UCL = 0.8534.

Since:

$$R = 1 - P_t \text{ (UCL)}$$

Then:

$R = 1 - (0.000004)\ (0.8534) = 0.9999966$, which is the predicted "reliability-in-use."

C Example No. 3

When nothing is known about the expected conditions in use, the "most severe condition in use" an item can withstand can be calculated:

Given:

Required "reliability-in-use" = 0.9999

Test condition used (i) = 10 G's

Number of items tested = 10

Number of failures obtained = none

Since:

$$R = 1 - P_t(UCL)$$

$$P_t = \frac{1 - R}{UCL}$$

From Table IX of reference 8, UCL for no failures in 10 trials equals 0.3085, for the two-sided 95% confidence level.

Then:

$$P_t = \frac{1 - 0.9999}{0.3085} = 0.000324$$

From Table 39 of reference 9, the "most severe condition in use" (m) for i = 10 and P_t = 0.000324 is found to be:

$$m = 2.6 \text{ G's}$$

With the given test result and test condition, this is the "most severe condition in use" under which the item will have 0.9999 reliability.

REFERENCES

1. Lusser, Robert, "Unreliability of Electronics - Cause and Cure," Redstone Arsenal, Huntsville, Ala., Nov. 1957.

2. Lusser, Robert, "Reliability Through Safety Margins," Redstone Arsenal, Huntsville, Ala., Oct. 1958.

3. Dixon and Massey, "Introduction to Statistical Analysis," McGraw-Hill Book Co., Inc., NYC, 1957, Second Edition.

4. Davies, O. L. (Editor), "Statistical Methods in Research and Production," Hafner Publishing Co., NYC, 1957, Third Edition.

5. Plackett and Burman, "The Design of Optimum Multifactorial Experiments," Biometrika, Vol 33, 1946, page 305.

6. "Fractional Factorial Experiments Designs for Factors at Two Levels," NBS Applied Math. Series No. 48.

7. Fisher, R. A., "Statistical Methods for Research Workers," Hafner Publishing Co., NYC, 1950.

8. Mainland, Herrera, and Sutcliffe, "Tables for Use With Binomial Samples," Department of Medical Statistics, NY University College of Medicine, NYC, 1956.

9. Pearson, E. S., and Hartley, H. O., "Biometrika Tables for Statisticians," Volume I, Cambridge University Press, Cambridge, England, First Edition, 1956.

MULTI-DIMENSIONAL STAIRCASE DESIGNS FOR RELIABILITY STUDIES

David R. Howes
U. S. Army Chemical Corps Engineering Command

This paper suggests the need for a sequential staircase procedure whereby a given contour of response could be traced experimentally without necessarily defining the entire response surface. Such a method would have important application in reliability studies, in design and engineering, etc. where the intent is to hold malfunctioning of some type at some predetermined level.

As an example, suppose that an artillery shell is to be filled with poison gas and closed with a burster tube (Figure 1, on the next page). It has been found that the leakage of these shells is affected by two variables, the interference of the burster tube (Variable A) and a variable B which is a structural characteristic of the shell body.

Although it may be possible to specify levels of A and B which will be satisfactory, it is also necessary to know the threshold of leakage in order to set manufacturing tolerances, filling procedures, etc. This involves the response surface generated in leakage in the A-B space.

Possible Methods

1. Factorial

It is possible to fit a response surface to the results of the experiment shown in Figure 2 (see next page) by well-known methods. The sample size, 1800, may seem excessive for accuracy which cannot exceed 2%.

2. Confined Factorial and Staircase Method

Staircase methods have been described which permit the isolation of percentage points of a problem with only one variable.[1,2,3] For a two variable case it would be possible to treat one variable by a staircase method, and the other factorially. (See Figure 3).

3. Multi-Dimensional Staircase Method

An extension of the staircase method to N variables is possible, although the methods have not been produced yet, since the theory doesn't exist. We would assume that a smooth response surface existed and that it would be possible, staircase-wise to follow some response contour on that surface say 90%, 93%, 99%, etc. Interaction of the variables is also a possibility.

The factorial approach may be inefficient here since it collects data not needed merely to trace a contour. It may also be wasteful of time, materials, and manpower, since the experiment cannot proceed until a large predetermined number of items are available. It should also be mentioned that it runs against the strong desire usually found

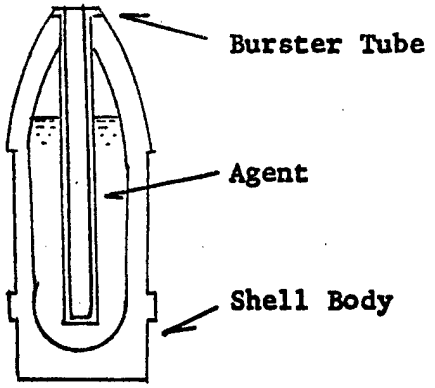


Figure 1.

Agent Filled Shell

Variable A Level	Number of Shells					
6	50	50	50	50	50	50
5	50	50	50	50	50	50
4	50	50	50	50	50	50
3	50	50	50	50	50	50
2	50	50	50	50	50	50
1	50	50	50	50	50	50
Levels	1	2	3	4	5	6

Variable B

Figure 2.

93% Points for Staircase Method

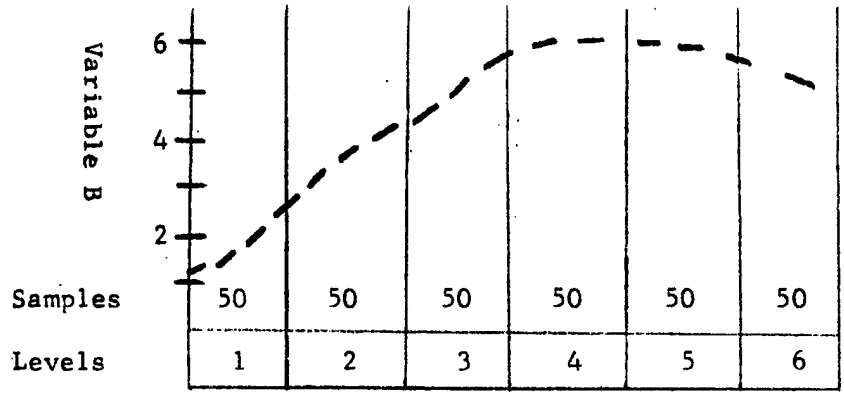


Figure 3.

X = leaker

0 = non-leaker

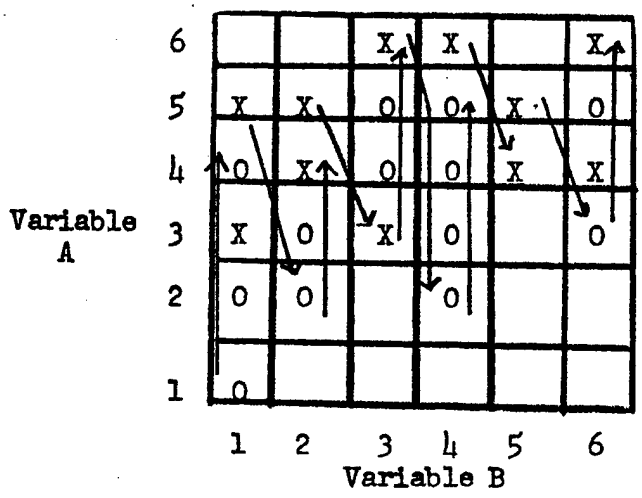


Figure 4.

among reliability engineers to try something, then try something else. Since the statistician must live with this, it would seem most desirable to adopt procedures which resemble to as great an extent feasible, those used by the reliability and test engineers. The sequential nature of the procedures has also the advantage of allowing experimentation to continue, while new experimental vehicles are being fabricated.

Example of a Multi-Dimensional Method:

In this procedure we might try a single sample at a time, and move over the response surface in accordance with certain rules based on the previous test result. (See Figure 4).

In this design, we proceed upward one step at a time in A in level No. 1 until two leakers have been found, then proceed to the next higher level of B at a level of A, one level below that on which the first leaker occurred. Then proceed upward in A until again two leakers have been found and proceed as before. Take the mid-point on A between the two leakers at each level of B as a point on the 50% leakage contour. Repeat this experiment as many times as necessary to get the desired precision of estimate. Instead of taking a single sample, we might take a sample of n and consider it "reliable" if no more than c leakers are found. This would, I suppose, lead to the tracing of percentage lines. Using a table of random numbers, I was able to get fairly good results in tracing a 10% contour over a bi-variate surface; good enough results to suggest the desirability of answering the following questions:

(1) What are the most effective rules to follow when traversing the response surface?

(2) What confidence can be placed in the results of K trials, using a sample size n in each trial, with an allowance of c defects as an estimate of the c/n fraction line?

(3) What method of computation or statistic would be used to obtain the c/n fraction line estimate from the data?

REFERENCES

1. National Defense Research Council Applied Mathematics Panel Report No. 101.1

2. Bureau of Ordnance, U. S. Navy, NAVORD Report No. 65-46, 21 March 46 - Anderson, McCarthy, & Tukey

3. Dixon, W. J. and Mood, A. M., "A Method of Obtaining and Analyzing Sensitivity Data" Journal of the American Statistical Association. Vol 43, No 241 (1948)

A PROPOSED RESEARCH PROGRAM FOR PROVIDING A QUANTITATIVE BASIS FOR PREVENTIVE MAINTENANCE POLICIES ON ORDNANCE EQUIPMENT

Walton M. Hancock
and
Randall E. Cline

This talk contains an outline of a proposed program which can be used to aid in the establishment of preventive maintenance policies. The program represents a general approach applicable to both existing systems and systems yet to become operational. It is anticipated that this approach should ultimately lead to the simplification of the maintenance of Ordnance equipment.

Since the Ordnance Corps has many different types of equipment, which vary both in complexity and density, no one PM policy can be applicable to all types of equipment. The research effort must therefore be oriented towards developing the proper general approach to the establishment of preventive maintenance policies for a variety of weapons systems. The applicability and usefulness of the approach, then, can be demonstrated by selecting a limited number of weapons systems and evaluating different preventive maintenance policies for them.

The talk is presented as follows: Part 1 contains a discussion of work done by others and comments on methods used to develop general solutions. Part 2 presents a general approach to the preventive maintenance problem, and Part 3 is a mathematical formulation which has been used to present, in a compact form, the ideas developed in the general approach. The fourth part contains a proposed program which will provide information for the evaluation of preventive maintenance policies on specific weapons systems in order to serve as examples of the application of this general approach to establishing preventive maintenance policies for other weapons systems.

1. CURRENT STATUS OF RESEARCH ACTIVITY CONCERNING MAINTENANCE POLICIES. An extensive library search and a number of visits have been made in order to take advantage of work performed by other research groups in the maintenance area.[1] One finds that there has been quite a bit of effort put into the specific details of establishing preventive maintenance policies for a particular type of equipment; there are, however, relatively few people that are concerned with a general approach.

The areas in which much work has been done include the military electronics field and the civilian trucking industry. An examination of the approaches used in these areas has proved quite helpful.

In the military electronics field, investigations have been made into the problems of reliability. The reliability of equipment is directly related to the amount and type of maintenance. Methods of analysis used in evaluating and improving reliability can also be used in developing maintenance programs. Briefly, the methods have been as follows:

---

1. A list of the most pertinent books and articles is given in the bibliography.

a. Rather than conduct broad scale studies, particular using units have been selected for detailed study.

b. Only new equipment or equipment put through major overhaul is issued to the units to be studied.

c. Detailed records are kept on the life history of the equipment. The amount of scheduled and unscheduled maintenance, the time required for repairs, basic causes of failure, parts usage and the time interval between failures are all carefully documented.

d. The reliability of the equipment is related to the mission that is to be performed, and the mean time to failure is frequently used in deriving an expression of reliability. Since the missions are usually expressed in terms of the number of hours of use per mission, then the probability that the equipment will perform the expected mission can be predicted.

e. Emphasis has been placed in the classification of the types of failures by basic components of the system such as by tube types, types of resistors, and capacitors. The reliability of the equipment is then expressed in terms of the reliability of its components.

Considerable success has been attained by the use of the above methods. The most notable of these have been realized by classifying the basic causes of failure for design purposes.

Since maintenance is one of the principal costs in the civilian trucking industry, effort has gone into the solution of their maintenance problems. The following represents the general methods used by the industry:

a. A detailed life history is kept on each vehicle. These data contain a record of all repairs, including parts usage and costs, and a record of all maintenance performed. Incidence of breakdowns, associated costs, and the mileage of the vehicle are recorded for such events in the life history.

b. By analysis of the life histories, norms are established for the expected life of each major component. Careful investigations are then made to determine the causes of failures which seem premature. Failures are also classified as to design deficiencies, improper maintenance or poor driving. Those that are in the design deficiency category are used to change the specifications on new equipment. Those caused by improper maintenance are used to modify the amount, kind, and time interval of scheduled maintenance. The failures caused by poor driving are analyzed for improved driver education.

c. The "cost of maintenance per mile" is also derived from the detailed life history. This is a control technique used by

administrative personnel to see if the total cost of maintenance is kept within prescribed limits.

Proper feedback of information and the proper analysis of this information is considered by most trucking firms to be an absolute necessity. This applies to commercial trucking fleets, bus fleets, truck and car rental fleets, and users of off-the-road equipment. The same approach prevails regardless of the actual use to which the equipment is put.

2. A GENERAL APPROACH TO THE MAINTENANCE PROBLEM. In order to develop a general approach that can be applied to all types of weapons systems, it is first necessary to classify weapons systems in such a way that their common characteristics as well as their differing characteristics are evident. Preventive maintenance policies will then be related to these characteristics. It appears that all weapons systems can be classified (for purposes of establishing maintenance policies) in terms of the following basic parameters:

a. Complexity

b. Density

c. Mission

Each of these three basic classifications is a vector quantity, or stated more simply, each can be described in terms of a number of factors. For instance, the complexity of a weapons system may be defined in terms of the crew requirements, the number of components, the total cost of the system, the average amount of time required to locate troubles, the ratio of time to check out the system compared to the time to complete a mission, etc. Similarly, density may be expressed in terms of geographical dispersion of equipment, travel time from support unit to supported units, total number of units in the field, etc. Missions can be defined in terms of the time equipment is required to be operable, the movements of operations which must be accomplished, the precision with which operations must be performed, etc.

Examples of the way these classifications are related to maintenance policies are as follows: Experience has shown that for electronic equipment an increase in complexity increases the maintenance requirements. The density of weapons systems affects the organization of maintenance crews and supporting test equipment. The mission also affects the type of maintenance. Since many weapons systems are required to perform a number of different missions, to achieve simplicity of maintenance at a minimum cost, the maintenance requirements may also vary. For example: trucks that are used on hard surface roads will require different maintenance than trucks used off the road. One trucking firm that was visited had different maintenance schedules for long distance vehicles than for local haul equipment, because the cost of a breakdown of a vehicle some distance from maintenance support was many times higher than for a vehicle used locally.

3. MATHEMATICAL FORMULATION OF A GENERAL APPROACH. Using the classification of weapons systems introduced in Part 2, a general approach to establishing preventive maintenance policies will now be considered. In the analysis of failure data for electronic equipment, the term reliability of a system or a component of a system has been generally used to denote the probability that a system or component will perform its required mission under given conditions for a specified operating time. The reliability of Ordnance equipment can be defined similarly. In developing a model which relates maintenance and reliability, the following assumptions are made:

a. Reliability is dependent upon the age of the equipment. For example, tanks, trucks, missiles, etc., tend to fail more frequently as the equipment becomes older.

b. The reliability is also dependent upon past usage of the equipment. For example, it is expected that the number of failures in trucks increase as the number of miles traveled increases. Similarly, the number of times a missile is checked is believed to affect the probability that the missile will fire.

Since both age and past usage are assumed to affect the reliability of equipment, it is useful to redefine reliability. Consequently, the following notation will be introduced. For any given system, let t = calendar age of the system, i.e., the number of years since the equipment was issued (new) to the user, and let $x(t)$ = usage of the system prior to time t, i.e., $x(t) = \int_0^t f(\tau)d\tau$, where $f(\tau)$ is some measure of usage.

The reliability of the system will be defined as follows:

The reliability of a system is the probability that the system will perform its required mission (which includes a specified operating time), given that the age of the system, t, and the past usage, $x(t)$, and the preventive maintenance policies are known. The reliability of a system will be designated by the symbol $r[t, x(t)]$.

It has been suggested that for wheeled and tracked vehicles, $x(t)$ = mileage, and for missiles, $x(t)$ = number of times certain checks have been made on the system.

Both scheduled and unscheduled maintenance are related to the reliability of a system. The purpose of maintenance is to increase $r[t,x(t)]$, and hence to maintain the reliability above some predetermined minimal level. The policies regarding scheduled and unscheduled maintenance are expected to be influenced by the complexity and by the density of the system. Visual inspections and operational check-out procedures are designed to ascertain whether the reliability of the system is above this predetermined level. Since a weapons system may be required to perform several different missions, we must consider all operations which the system may be required to perform and the associated performance times. Classify these missions (that is operation-time combinations) in groups in such a way that all missions in any given group are roughly equivalent

in terms of requirements on the system. Such a group of missions will be called a task. Now order these tasks in such a way that if the system can perform any given task, then it can also perform all simpler tasks. Designate these tasks by $T_1, \ldots, T_\ell$, $(\ell \geq 2)$, where T_1 indicates the simplest task and T_ℓ corresponds to the most difficult task. Having specified a given task, say T_k, there exists a corresponding probability that the system can perform this task. Denote this probability by $r_k[t,x(t)]$.

Then for fixed t and x(t), and all $k = 2,\ldots, \ell$,

$$r_k[t,x(t)] \geq r_{k+1}[t,x(t)] .$$

Now for each task T_k and any given maintenance policy, there exists a surface $r_k[t,x(t)]$ which represents the reliability of the system relative to T_k. Such a surface is illustrated in Figure 1.

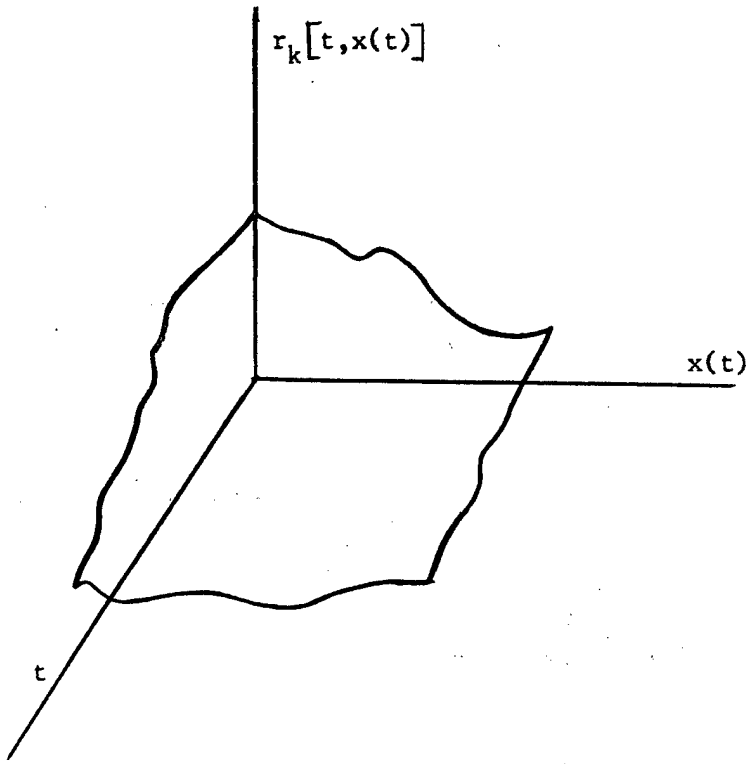


Figure 1

Theoretically, a maintenance policy should be designed in such a way that for any value of t, $0 \leq t \leq t'$, where t' is that time at which the system is discarded, and for any usage x(t), the reliability of the system should be maintained in such a way that for some α, $0 < \alpha \leq 1$,

$$r_k[t, x(t)] \geq \alpha \tag{1}$$

for some task T_k. Since it is assumed that the reliability of a system decreases with both time and usage, then maintenance must be performed in an attempt to satisfy equation (1). The physical situation is illustrated in Figure 2 in which the lower surface represents the actual system reliability and the upper surface represents the desired goal of maintenance. It is to be noted that $r_k[t,x(t)]$ may actually exceed α. Conversely, the effect of inappropriate design on the system may be such that $r_k[t,x(t)]$ never attains the desired goal. Finally, it is observed that the goal may not be constant over the entire expected life of the system and may be lowered as equipment is phased out.

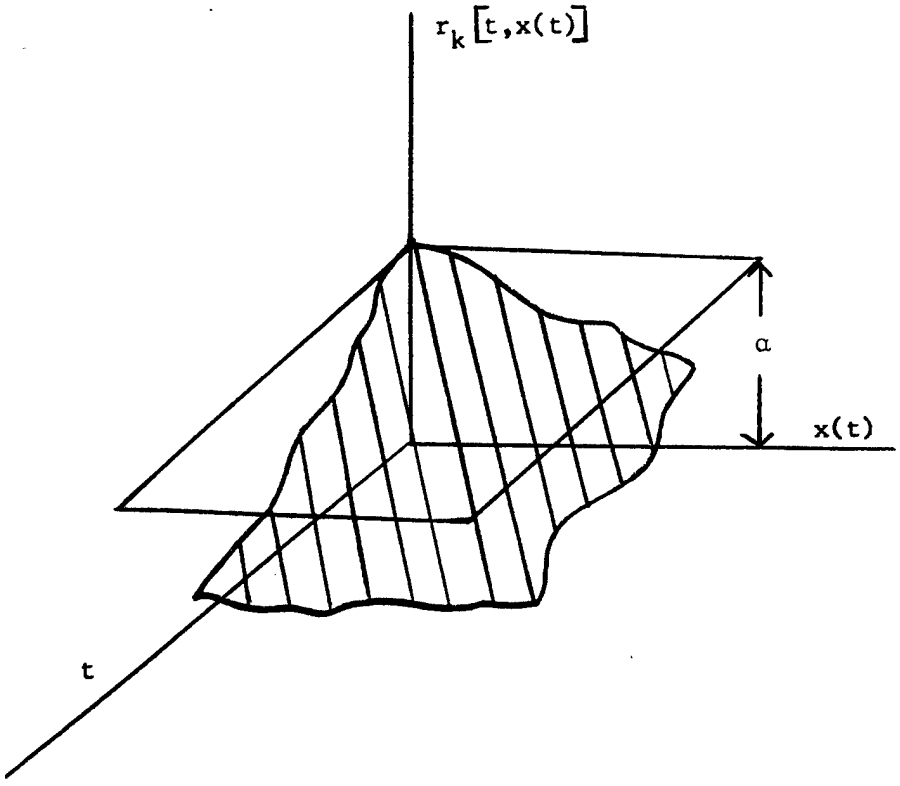


Figure 2

The life history of a system can be represented by a curve in the t,x(t) plane. Associated with this curve is the corresponding reliability. Such a curve is shown in Figure 3.

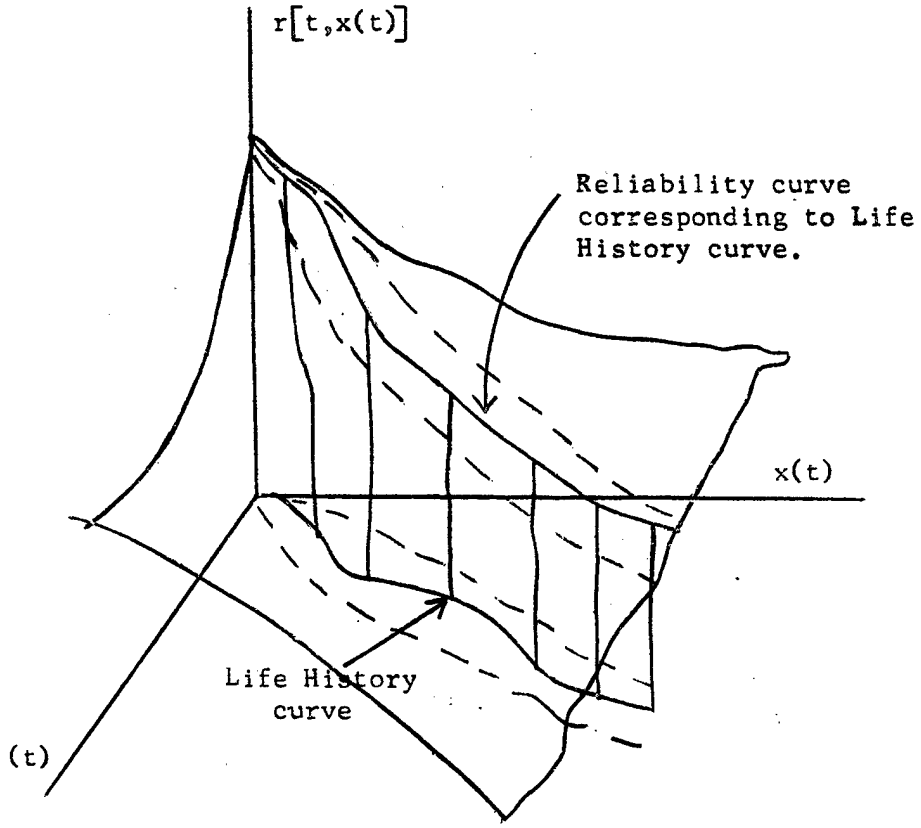


Figure 3

Observe now that since a collection of similar weapons systems in an Army unit will not be used in identical amounts, then observations taken from a particular unit will form a wedge as illustrated by the broken lines in Figure 3.

As mentioned above, maintenance of any type is intended to increase the reliability of the system. Thus, for any maintenance performed at a given point $[t,x(t)]$, let:

$$\Delta r_k [t,x(t)]$$

denote the change in $r_k[t,x(t)]$ obtained by performing the maintenance. The amount of this change is a random variable dependent upon the type of maintenance performed, the level of skill of the technician performing it, and the tools or test equipment available to him. Graphically, this may be illustrated as in Figure 4 for a particular system having maintenance performed at points $[t,x(t)]$, $[t_2,x(t_2)]$, ..., $[t_n,x(t_n)]$, where the jumps in the curve indicate the corresponding changes in reliability resulting from the maintenance.

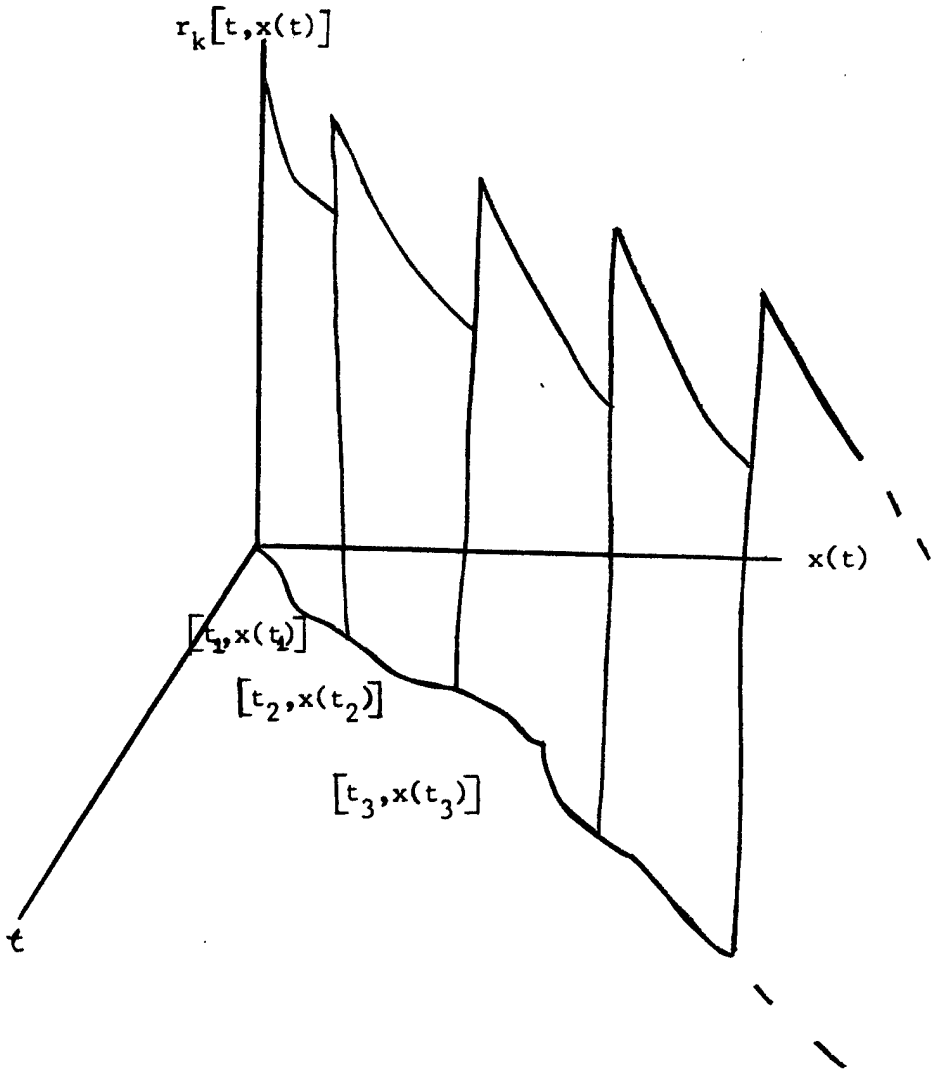


Figure 4

It is to be observed that if not only $r_k[t,x(t)]$ and $\Delta r_k[t,x(t)]$ are known, but also various costs of maintenance associated with $\Delta r_k[t,x(t)]$, then measures can be developed which relate costs of maintaining equipment to the expected loss from having equipment inoperative. Consequently, to use this approach for evaluating the effects of various preventive maintenance policies in terms of the change of the reliability of the system, it will be necessary to develop techniques for estimating the forms of $r_k[t,x(t)]$ and $\Delta r_k[t,x(t)]$ for various tasks and maintenance practices.

To aid in the estimation of these functions, it is useful to consider the effects of various components of the system on overall system performance. Now any given system can be represented as a collection of major components (or subsystems). Associated with each component is a

corresponding reliability surface again a function of t and x(t). Assuming these major components are serially connected and statistically independent, then at any point $t,x(t)$,

$$r_k\left[t,x(t)\right] = \prod_{j=1}^{n} r_k^{(j)}\left[t,x(t)\right]$$

where there are n major components with reliabilities $r_k^{(j)}\left[t,x(t)\right]$, $(j=1,\ldots,m)$. By estimating not only $r_k\left[t,x(t)\right]$ and $\Delta r_k\left[t,x(t)\right]$, but also the corresponding quantities for major components, those components requiring the most maintenance will be apparent. Thus, concurrent with the collection of data to be used in estimating the overall system reliability, data on various components will also be collected. It is to be observed that this additional component analysis is essential for systems in which major components have been replaced.

Continued work on the general formulation will be directed toward relating the function $r_k\left[t,x(t)\right]$ and $\Delta r_k\left[t,x(t)\right]$ both to the organization of maintenance in terms of costs, skills, tools, etc., and to the complexity, density and task classifications of weapons systems.

4. PROPOSED PROGRAM FOR FIELD OBSERVATION AND DATA ANALYSIS. To obtain the information required to develop the curves and surfaces discussed in Part 3, a field observation program will be necessary. It is necessary that at least two, and preferably three, weapons systems be selected that represent different points in the complexity-density range. Since life histories are to be collected, new weapons systems should be selected whenever practicable. The specific weapons systems that are to be studied have not been chosen at the present time. A wheeled or tracked vehicle, a missile, and possibly a hand weapon will probably be selected. In selecting the weapons system and the units to be observed, proper consideration also must be given to the missions that are being performed.

Since a main purpose of the research is to quantitatively evaluate the effect of preventive maintenance practices to permit determination of the proper amount of scheduled maintenance that should be performed, it will be necessary, after an initial observation period, to slightly modify the existing maintenance practices and observe their effects.

Specifically, we will attempt to collect the following information:

A. Scheduled Maintenance

For each scheduled maintenance or operational check performed, the following information is desired:

1. Type of scheduled maintenance

2. Frequency of scheduled maintenance

3. Equipment usage between scheduled maintenance periods.

4. Deficiencies found during scheduled maintenance.

5. Parts replaced during scheduled maintenance.

6. Time to repair deficiencies found during scheduled maintenance periods.

7. Time to perform scheduled maintenance.

8. The echelons that perform the scheduled maintenance.

B. Unscheduled Maintenance

For each failure requiring unscheduled maintenance the following information is desired:

1. Frequency of failure

2. Basic cause of failure

3. Elapsed time since specific scheduled maintenance.

4. Usage since specific scheduled maintenance.

5. Parts needed for repair of failure

6. Parts available

7. Time to repair failure

8. Echelon performing the repair

C. Maintenance Organization

For the unit being observed the following information is desired:

1. Skills and equipment available at using unit and supporting units.

2. Inspection criteria at each echelon

3. Work load at each echelon.

We plan to initially place technically qualified field engineers on a full time basis with the units to assist user personnel in recording the above information. It is hoped that after approximately three to six months the cooperating units will be able to provide the necessary information, and the field staff will be required only to monitor the data collection program on a part-time basis. This would also free the field staff to initiate a similar observation program with an additional using unit.

As discussed in Part 3, the data collection program will be initiated concurrently with the continuation of the development of the general approach. As the program proceeds, the level, type and method of data collection may, of course, require modification. The initial information obtained will aid in selecting the most pertinent of many possible characteristics of weapons systems for first consideration in the mathematical formulation, and in defining the groups of missions required of the different types of systems studied. Such effort will be needed to further refine the general formulation to insure that the results will have practical significance to the particular weapons systems under consideration.

BIBLIOGRAPHY FOR APPENDIX III

1. "Effects of Maintenance on System Reliability," ARINC Research Corporation, Publication No. 101-16-144, Sept., 1959.

2. "An Analysis of the Effects of Maintenance on Parts Replacements," Ralph L. Madison, AERO. Radio Inc., Publication No. 112, Jan. 6, 1958.

3. "A Survey of the Current Status of the Electronic Reliability Program," R. R. Carhart, The Rand Corporation, RM 1131, Aug., 1953.

4. "Maximizing Expected Machine Up Time," Jay B. Heyne, Systems Development Laboratory, Hughes Aircraft Co., July, 1959.

5. "Evaluation and Prediction of Circuit Performance by Statistical Techniques," J. Marini, H. Brown, R. Williams, ARINC Inc., Publication No. 113, Feb., 1958.

6. "The Condition of 60 Farm Tractors Point Ways to Improve Maintenance," J. A. Weber, Society of Automotive Engineers, Paper No. 181, Sept., 1957.

7. "Problems Relating to Systems Checkout and Final Acceptance of Production Missiles," M. A. Stephens, Society of Automotive Engineers, Paper No. 48B, April, 1958.

8. "A Bibliography of Life Testing and Related Topics," William Mendenhall, Biometrika, Vol. 45, Parts 3 and 4, pp. 521-543, Dec., 1958.

9. "Maintenance Management--Operational Readiness Plan for the Redstone Missile System," Lt. Col. Brian O. Montgomery, Redstone Arsenal, Alabama, Sept., 1958.

10. "Optimum Component Redundacy for Maximum System Reliability," Robert Gordon, Operations Research Society Journal, Volume 5, No. 2, April, 1957. (pp. 229-243).

11. "Reliability of Military Electronic Equipment," The Advisory Group on Reliability of Electronic Equipment, Office of the Assistant Secretary of Defense (Research and Engineering), 4 June, 1957.

12. "Queues, Inventories, and Maintenance," Philip M. Morse, John Wiley and Sons, 1957.

13. "Reliability Factors for Ground Electronic Equipment," Henney, Lopatin, Zimmer, Poler, and Naresky, McGraw Hill, New York, 1956.

14. "A Good Preventive Maintenance Program," SAE Journal, Vol. 65, p. 88, July, 1957.

15. "Automatic Transmissions Make Some Truck Operator's Happy, Others Sad," SAE Journal, Vol. 65 pp. 56-58, Spetember, 1957.

16. "Simplicity - Key to Military Design," SAE Journal, Vol. 67, pp. 34 and 35, July, 1959.

17. "Air Force Model Shop Sets New PM Pace," Commercial Car Journal, Vol. 90, pp. 74-75, December, 1955.

18. "Maintenance Yardsticks Show Bus PM Efficiency," Commercial Car Journal, R. W. Ziffle, Vol. 90, pp. 68, 128-132, December, 1955.

19. "Systems Checkout Insure Missile Reliability," M. A. Stevens, SAE Journal, Vol. 66, pp. 80-81, August, 1958.

20. "Do Road Calls Measure PM Efficiency?" Commercial Car Journal, Vol. 90, pp. 71 and 186, September, 1955.

A STATISTICAL ANALYSIS OF VARIOUS PARAMETERS AFFECTING THE BURNING CHARACTERISTICS OF FLARE SYSTEMS

Bossie Jackson
Pyrotechnics Chemical Research Section
Pyrotechnics Laboratory
Picatinny Arsenal

The development of pyrotechnic flare compositions involves the investigation of a number of different variables. From the many investigations conducted previously numerous hypotheses were formed concerning the relationship of such variables as candlepower and burning rate with flare case coating, loading pressure, and the particle size of the fuel employed in the system. A better knowledge of the basic factors governing the burning characteristics of solid mixtures was desired. It was expected that the results of this investigation would tend to substantiate the various hypotheses.

Previous data tends to show that candlepower and burning rate are depended upon particle size, loading pressure, and are not affected by flare case coating. The analysis of these relationships was based on data obtained from this study using two statistical methods, the test for least significant differences, and in particular the analysis of variance. The experimental design for these studies is given in Figure 1. (See next page.) This configuration was used for four (4) levels of magnesium, M_1, M_2, M_3, and M_4. The flare case coatings are shown by C_1, C_2, C_3, and C_4. Seven (7) levels of loading pressure were studied ranging from P_1 to P_7. Five samples were utilized for each combination of pressure, case coating, and magnesium particle size.

A standard flare composition (Table I) [Tables can be found at the end of this article] was used for this experiment. This composition contains 48% magnesium, 42% sodium nitrate, 2% polyvinyl chloride, and 8% Laminac resin. Flare compositions are consolidated in a variety of cylindrical cases under a specified pressure to obtain a cigarette-type propogating composition rather than one which flashes or explodes. The use of self-hardening polyester resins in flare compositions eliminated the need for consolidating them at very high pressures. Most flare compositions are presently loaded at pressures ranging from 4000 pounds per square inch (psi) to 10,000 psi.

Standard flare compositions each containing different mesh sizes of magnesium were evaluated in this study. These magnesium granulations together with their average particle diameter are given in Table II. The mesh sizes of magnesium are 20/50, 30/50, 50/100, and 100/200 with particle sizes varying from 437 microns to 110 microns. The compositions were consolidated at 2000, 4000, 7000, 10,000, 15,000, 20,000 and 25,000 pounds per square inch.

The effect of loading pressure on candlepower can be observed in Figure II. The mean candlepower values vary from 201,000 to 223,000 which approximates an eleven (11) percent change. It is apparent from

this graph that a definite trend exists with candlepower increasing with increased loading pressure. Table III summarizes the luminous intensity values observed at each pressure level. The least significant difference value at the ninety-five (95) percent confidence level is also given for these values. Despite the apparent trend of candlepower values, it will be observed that the difference in light output from 4,000 psi to 25,000 psi are not large enough to be significant based on the least significant difference value. This conclusion also holds for the candlepower at 2,000 and 4,000 pounds per square inch. However, the differences between the values at 2,000 psi and those at 7,000 psi and above are large enough to be termed significant. The appearance of this trend may be attributed to the relationship between porosity and heat conduction. A more porous column will conduct heat at a slower rate as a result of the air pocket acting as insulators giving slow burning rates and low candlepower. Conversely, the less porous column will conduct heat at a faster rate giving higher candlepower and burning rates.

Higher candlepower values are obtained from smaller particle diameters of magnesium as shown in Figure III. Candlepower plotted as a function of average particle diameter decreased with increasing particle size. The higher candlepower value obtained for the 168 micron magnesium compared to the finer 110 micron magnesium may be attributed to the distributional effect of particle size. As evidenced by the low average particle diameter, the 50/100 mesh magnesium contained a large percentage of fines which placed the material in the upper range of the finer 100/200 mesh magnesium which may account for the higher light output. The candlepower values for each mesh size magnesium are tabulated in Table IV. It is immediately apparent that these intensity values are significantly different from each other on application of the least significant difference value. It is also observed that the 50/100 mesh magnesium gave significantly higher candlepower than the 100/200 mesh fuel. This result reflects the necessity for reducing the tolerance limits of particle size for the different mesh sizes of magnesium. It is believed that data can be accumulated from studies conducted previously to show that candlepower definitely decreases with increasing particle size.

It was previously mentioned that any of the investigators in the field of pyrotechnics believed that candlepower was unaffected by case coatings. This was verified by the results obtained from this experiment as shown in Table V. The flare case coatings studied were Amberlac 292, Laminac resin 4116, Polyethylene 617, and paraffin wax. The candlepower values vary from 210 to 218 and are essentially the same based on the least significant difference value of 8.8. Flare case coatings are especially necessitated with compositions containing self-hardening resins. Since these resins undergo considerable shrinkage on curing, voids and air pockets are created as a result of the composition separating from the flare case wall. Such a condition gives rise to possible detonations or increased burning rates.

Just as candlepower shows an insignificant trend resulting from increased loading pressure, burning rate values show a parallel trend. Figure IV illustrates this trend as the loading pressure is increased

from 2,000 psi to 25,000 psi. The burning rate values show a trend towards reduction with increasing loading pressure. Table VI tabulates the mean burning rate values obtained at each pressure level. It also summarizes that the differences in burning rates are not large enough to be significant. Based on this method of analysis, it can be concluded that burning rate is not affected by increasing loading pressure. As a result of the oppositely parallel trends shown by candlepower and burning rate, it can be hypothesized that these two variables are interrelated. This hypothesis is corroborated and borne out when considering the effect of particle size of magnesium on burning rate.

Figure V shows burning rate plotted as a function of average particle diameter. It can be seen that burning rate decreases with increasing average particle size. As shown on Table VIII the differences between these values were found to be significant based on the least significant difference value of 0.05. It was previously observed that a corresponding effect was obtained with candlepower values, except in the result for 50.100 mesh magnesium. To further complicate the picture it was observed that significant differences in burning rates existed for the various flare case coatings. The results are given in Table VIII. The burning rates vary from a slow 3.61 inches per minute for the polyethylene to a fast 4.52 inches per minute for paraffin wax with Amberlac and Laminac resins yielding values in the middle. These differences based on the test for least significant difference indicate that the minimum and maximum values here are significantly different from the intermediate ones which are essentially the same. The significant effects of flare case coatings are undoubtedly due to their variation in binding strength, rate of thermal degradation, and end-products produced on combustion.

Burning rates of pyrotechnic compositions are also derived from the weight composition undergoing combustion per unit time. Figure VI illustrates the effect of loading pressure on the grams of composition per second from 2,000 pounds per square inch to 25,000 pounds per square inch. It is shown that grams of composition per second tends to increase with increasing loading pressure. The change in pressed density from 2,000 psi to 25,000 psi approximated twenty (20) percent. As shown in Table IX the differences in grams of composition per second are not significant based on the least significant difference value of 0.60. It was previously observed that the linear burning rate was not significantly affected by loading pressure.

In direct contrast to the above result, it was determined that the average particle diameter of magnesium had a significant effect on the weight of composition consumed per unit time. Based on the previous effects of particle size on both candlepower and burning rate this result could be anticipated. Figure VII shows grams of composition per second as a function of average particle diameter. It is observed that the number of grams burned per unit time varies inversely as the average particle diameter. The mean burning rate values tabulated in Table X are shown to be significantly different from each other as a result of the test for least significant difference.

By observing Table XI, it can be seen that flare case coatings do not significantly affect the grams of composition per second. This conclusion results in the fact that the average particle diameter of magnesium is the only parameter that significantly affects candlepower, burning rate (inches per minute), and burning rate (gram per second). The only other parameter contributing a significant effect was flare case coating on the linear burning rate.

Summarizing the results previously discussed, the analysis of variance table for candlepower is observed (Table XIII). This table gives the main effects, first and second order interactions of the parameters under consideration. Where the calculated F-ratio exceeds the critical F-ratio the effect of the parameter is considered significant. The main effects of magnesium particle size and case coatings are in accord with the results based on the test for least significant difference. The main effect from loading pressure observed here is significant as opposed to its insignificant effect based on the least significant difference value. Considering that such large changes in loading pressure (92%) results in very small changes in light output, it may indicate that the increasing trend of candlepower is insignificant. The first order interactions of magnesium loading pressure and case coating-loading pressure show a significant effect on candlepower, while magnesuim-case coating is insigificant. Evidently, the flare case coating cancels out the effect of the magnesium particle size. The second order interactions of these parameters are shown to be insignificant.

Table VIII outlines the analysis of variance table for burning rate (inches)per minute). It is observed that main effects resulting from magnesium and case coating are significant and loading pressure insignificant. This corroborates the results based on the test for least significant difference. The first order interactions of magnesium-case coating and magnesium-loading pressure are shown to be significant, while case coating-loading pressure is insignificant. The second order interactions are significant.

The results of the analysis of variance for burning rate (grams per second) is given in Table XIV. Except for the fact that flare case coating is shown to be insignificant here, the results parallel exactly those obtained for the analysis of variance on the linear burning rate. This table also substantiates the results from the test for least significant difference.

TABLE I

COMPOSITIONS EVALUATED

Ingredients	Percent by Weight			
Magnesium, Atomized, 20/50, 437 microns	48			
Magnesium, Atomized, 30/50, 322 microns		48		
Magnesium, Atomized, 50/100, 168 microns			48	
Magnesium, Atomized, 100/200, 110 microns				48
Sodium Nitrate, 34 microns	42	42	42	42
Polyvinyl Chloride, 27 microns	2	2	2	2
Laminac Resin 4116*	8	8	8	8

* Laminac Resin 4116 - 98.5%
  Lupersol ddm - 1.0%
  Nuodex - 0.5%

TABLE II

MAGNESIUM GRANULATIONS

Mesh Size	Tapped Density, gm./cc.	Average Particle Diameter, Microns
20/50	1.07	437
30/50	1.08	322
50/100	1.14	168
100/200	1.08	110

TABLE III

TESTS FOR LEAST SIGNIFICANT DIFFERENCE
OF LOADING PRESSURE VS. AVERAGE CANDLEPOWER

Least Significant Difference - 13.8

Level of Confidence, % - 95.0

Loading Pressure Psi	Average Candlepower
2,000	201.0
4,000	213.0
7,000	215.0
10,000	217.0
15,000	218.0
20,000	222.0
25,000	223.0

TABLE IV

TESTS FOR LEAST SIGNIFICANT DIFFERENCE
OF MAGNESIUM MESH SIZE VS. AVERAGE CANDLEPOWER

Least Significant Difference - 6.6

Level of Confidence, % - 95.0

Magnesium Mesh Size	Trapped Density gms/cc	Average Particle Size, Microns	Average Candlepower
20/50	1.07	437	130.0
30/50	1.08	322	154.0
50/100	1.14	168	293.0
100/200	1.08	110	285.0

TABLE V

TESTS FOR LEAST SIGNIFICANT DIFFERENCE
OF FLARE CASE COATING VS. AVERAGE CANDLEPOWER

Least Significant Difference - 8.8

Level of Confidence, % - 95.0

Flare Case Coating	Average Candlepower
Amberlac Resin 292	218.0
Laminac Resin 4116	215.0
Polyethylene 617	218.0
Paraffin Wax	210.0

TABLE VI

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.61

Level of Confidence, % - 95.00

Loading Pressure, Psi	Average Burning Rate, Inches Per Minute
2,000	4.41
4,000	4.54
7,000	4.29
10,000	4.23
15,000	4.03
20,000	4.17
25,000	4.06

TABLE VII

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.05

Level of Confidence, % - 95.00

Magnesium Mesh Size	Average Burning Rate, Inches Per Minute
20/50	2.62
30/50	3.01
50/100	5.66
100/200	5.84

TABLE VIII

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.14

Level of Confidence, % - 95.00

Flare Case Coating	Average Burning Rate, Inches Per Minute
Amberlac Resin 292	4.13
Laminac Resin 4116	4.15
Polyethylene 617	3.61
Paraffin Wax	4.52

TABLE IX

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.60

Level of Confidence, % - 95.00

Loading Pressure Psi	Burning Rate Grams Composition/second
2,000	5.06
4,000	5.37
7,000	5.38
10,000	5.49
15,000	5.48
20,000	5.58
25,000	5.57

TABLE X

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.06

Level of Confidence, % - 95.00

Magnesium Mesh Size	Burning Rate Grams Composition per second
20/50	2.41
30/50	2.82
50/100	6.11
100/200	6.33

TABLE XI

TESTS FOR LEAST SIGNIFICANT DIFFERENCE

Least Significant Difference - 0.23

Level of Confidence, % - 95.00

Flare Case Coating	Burning Rate Grams Composition/second
Amberlac Resin 292	4.36
Laminac Resin 4116	4.32
Polyethylene 617	4.50
Paraffin Wax	4.50

TABLE XII

ANALYSIS OF VARIANCE TABLE FOR CANDLEPOWER VALUES

ANALYSIS OF VARIANCE

Source Main Effects	Sum of Squares	Degree of Freedom	Mean Squares	F - Ratios Calculated	Critical F - Ratios 95% Level of Confidence
Magnesium (M)	3,076,232.87	3	1,025,410.956	2220	2.60
Pressure (P)	26,818.30	6	4,469.716	5.18	2.66
Coating (C)	5,592.08	3	1,864.0266	2.29	3.86
Interactions, First Order					
M + C	7,334.05	9	814.8944	1.77	1.88
M + P	15,535.53	18	863.085	1.88	1.57
C + P	9,805.12	18	544.7289	1.90	1.66
Interactions, Second Order					
M + P + C	15,415.97	54	285.6661	<1	1.32
Residual Error	206,239.40	448	460.3558		
Total Sum of Squares	3,362,972.94	559			
Pooled Variance	460,355,800				
Standard Deviation	$\sqrt{460,355,800} = 21,455$				

TABLE XIII

ANALYSIS OF VARIANCE TABLE FOR BURNING RATE (INCHES PER MINUTE) VALUES

BURNING RATE, INCHES PER MINUTE

Source	Sum of Squares	Degrees of Freedom	Mean Square	F - Ratios Calculated	Critical F - Ratios 95% Level of Confidence
Main Effects					
Magnesium (M)	12,718,235	3	4,272,745	19,600	2.60
Pressure (P)	160,662	6	26,777	1.94	2.66
Coating (C)	61,301	3	20,651	15.3	3.86
Interactions First Order					
M + C	12,163	9	1,351	6.2	1.88
M + P	248,338	18	13,797	63.1	1.57
C + P	30,731	18	1,707	1.3	1.66
Interactions Second Order					
M + P + C	70,522	54	1,306	6.0	1.32
Residual Error	97,992	448	219		
Total Sum of Squares	13,399,945	559			
Pooled Variance	0.0219				
Standard Deviation	$\sqrt{0.0219} = 0.15$				

TABLE VIV

ANALYSIS OF VARIANCE TABLE FOR BURNING RATE (GRAMS PER SECOND) VALUES

Analysis of Variance — Burning Rate, Grams of Composition/Sec.

Source	Sum of Squares	Degree of Freedom	Mean Square	F - Ratio Calculated	Critical F - Ratio 95% Level of Confidence
Main Effects					
Magnesium (M)	18,333,445	3	6,111,148	176,000	2.60
Pressure (P)	150,184	6	25,031	1.89	2.66
Coating (C)	35,199	3	11,733	2.13	3.86
Interactions, First Order					
M + C	34,800	9	3,866	11.2	1.88
M + P	238,236	18	13,235	38.4	1.57
C + P	24,573	18	1,363	1.07	1.66
Interaction, Second Order					
M + P + C	68,748	54	1,273	369	1.32
Residual Error	154,606	448	345		
Total Sum of Squares	19,039,724	559			
Pooled Variance	0.0345				
Standard Deviation	$\sqrt{0.0345} = 0.19$				

Figure I

EXPERIMENTAL DESIGN

	M_1				M_2				M_3				M_4			
	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
P_1																
P_2																
P_3																
P_4																
P_5																
P_6																
P_7																

Figure II

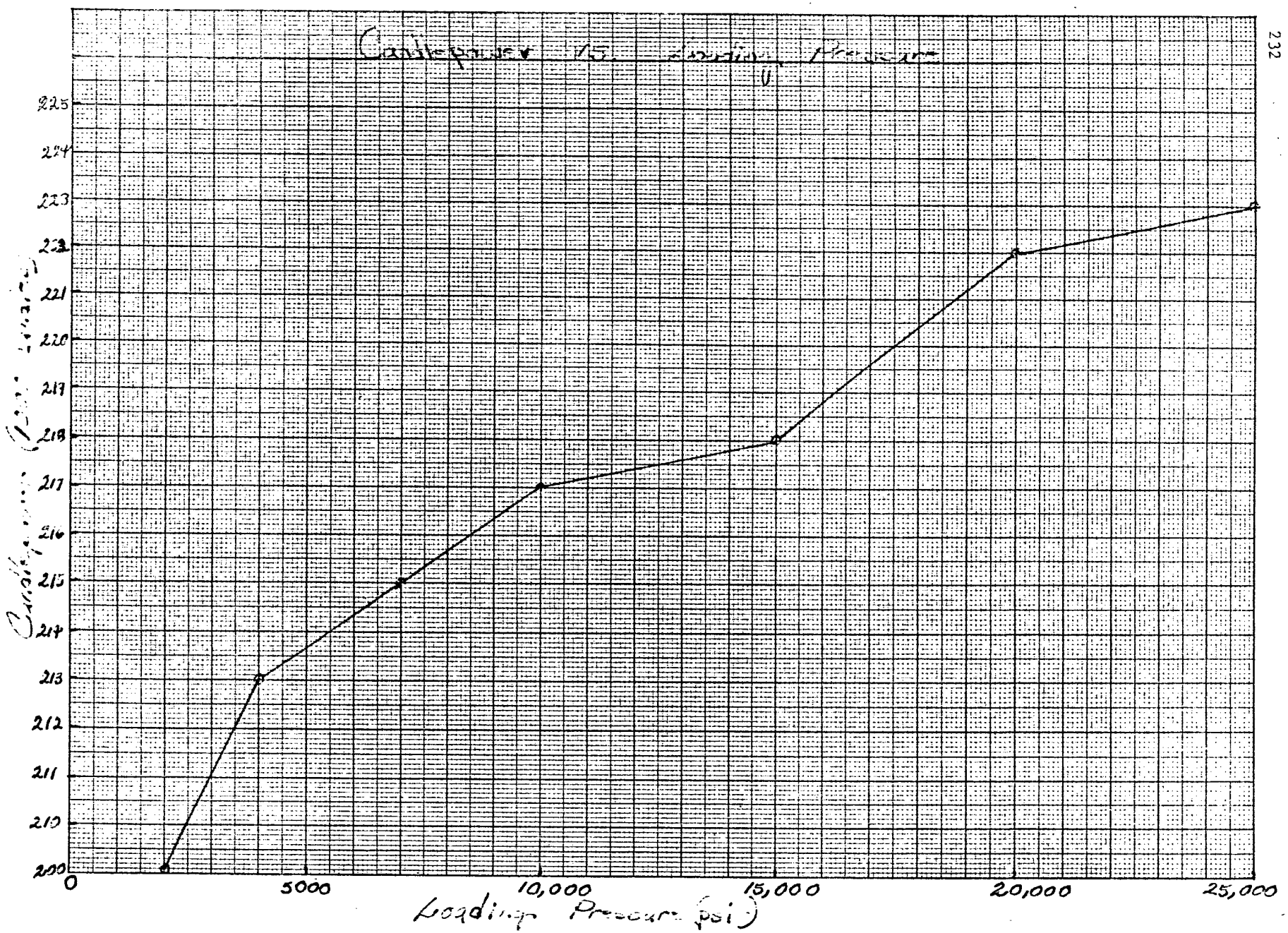
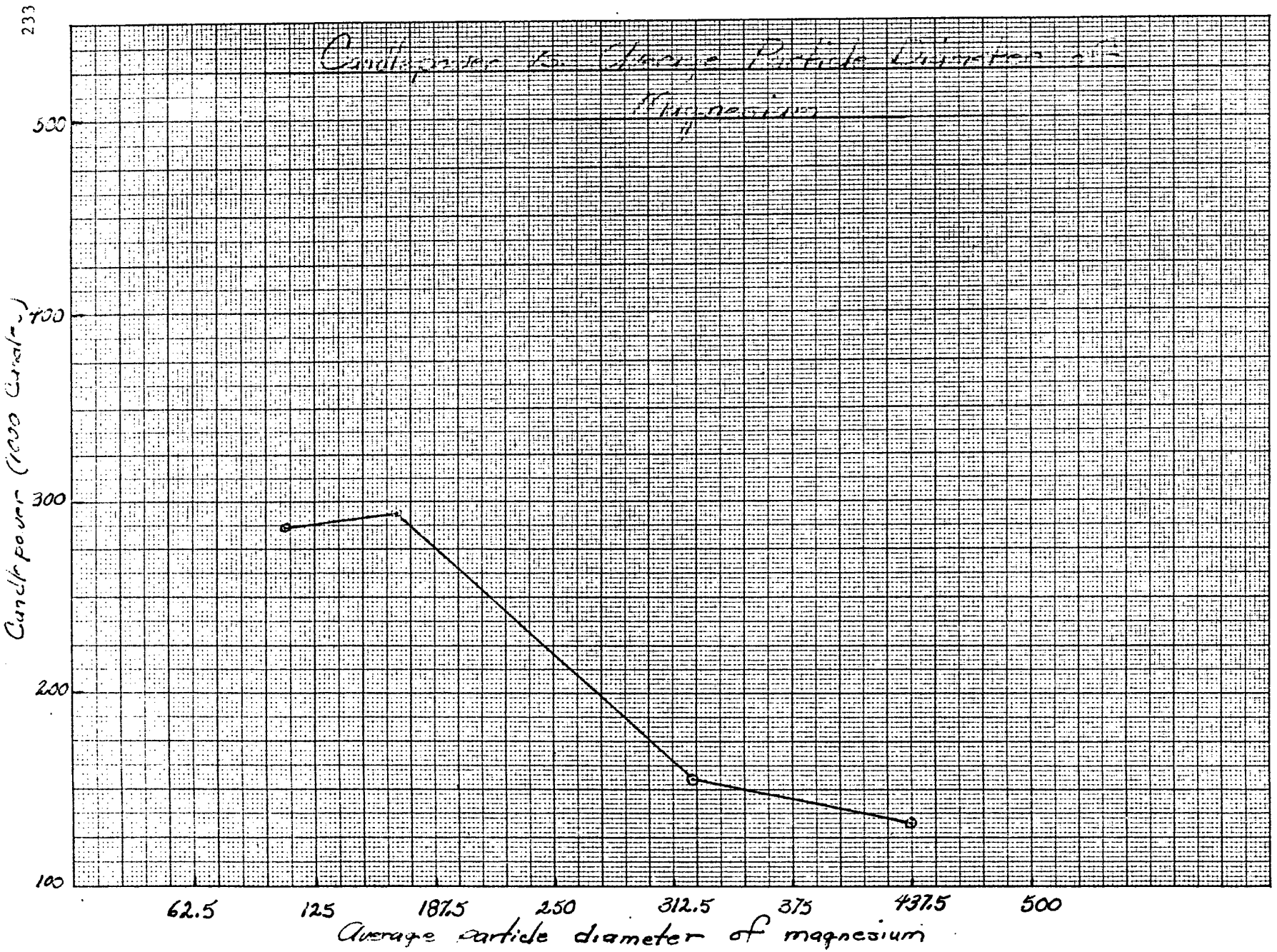

232
225
224
223
222
221
220
219
218
217
216
215
214
213
212
211
210
200
0
5000
10,000
15,000
20,000
25,000
Loading Pressure (psi)

Figure III


233
Candlepower (1000 Candle)
500
400
300
200
100
62.5
125
187.5
250
312.5
375
437.5
500
Average particle diameter of magnesium

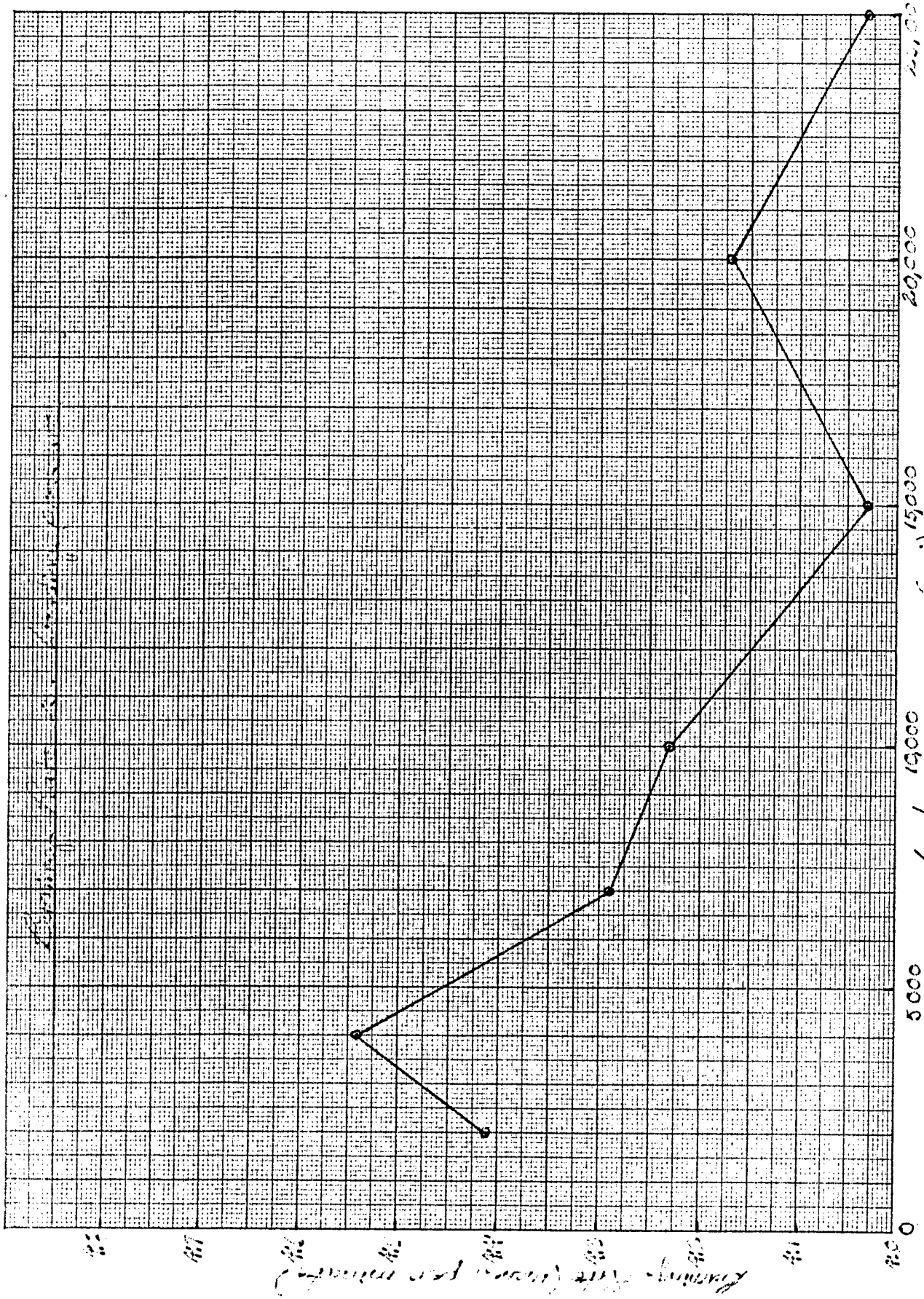


Figure IV

Figure V

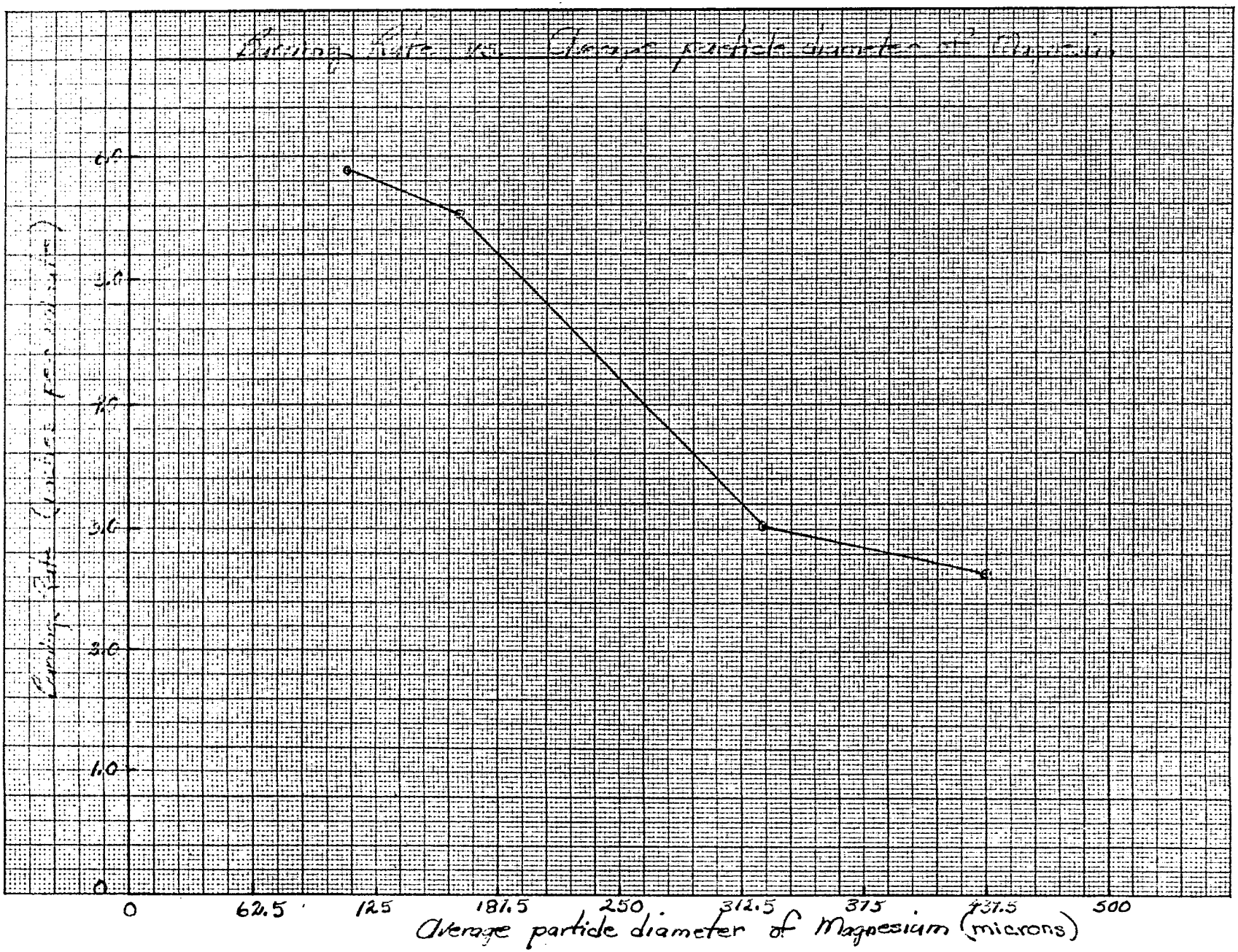

6.0
5.0
4.0
3.0
2.0
1.0
0
0
62.5
125
181.5
250
312.5
375
437.5
500
Average particle diameter of Magnesium (microns)

Figure VI

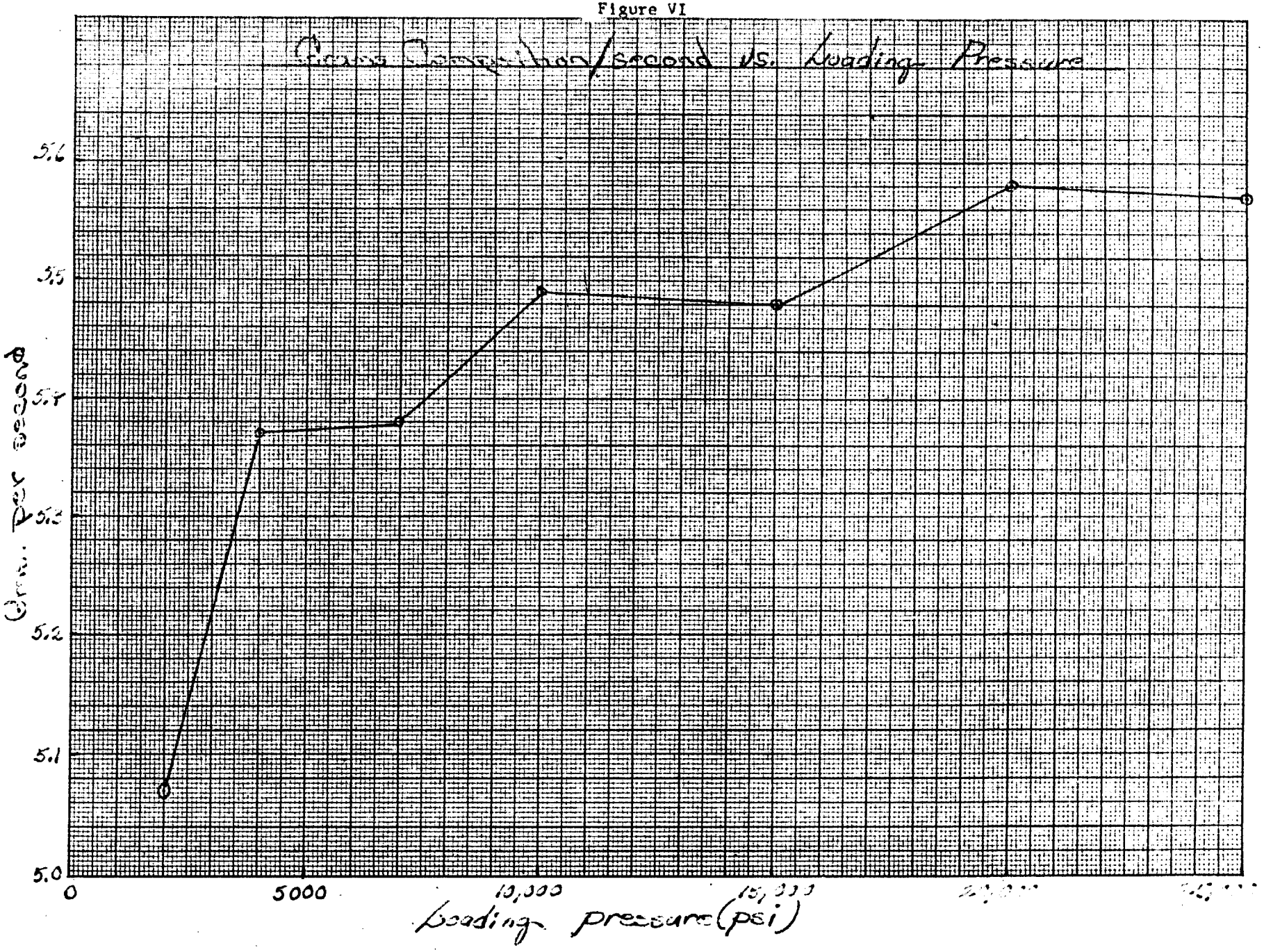
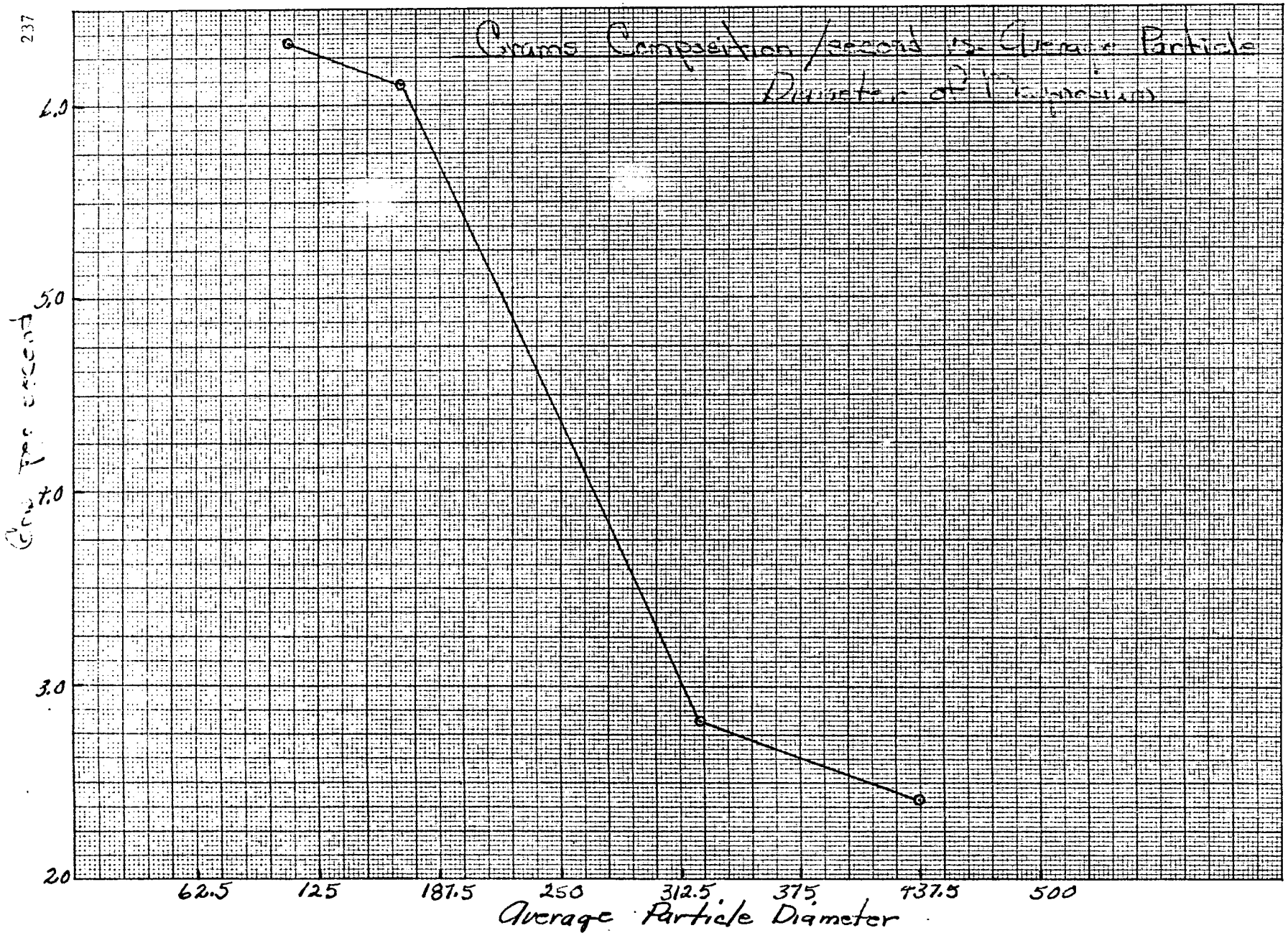

Grams Compression/Second vs. Loading Pressure
5.6
5.5
5.4
5.3
5.2
5.1
5.0
Grams per second
0
5000
10,000
15,000
Loading pressure (psi)

Figure VII


237
Grams Composition/Second vs. Chemical Particle
Diameter of Tobacco Stem
Grams per second
4.0
5.0
7.0
3.0
20
62.5
125
187.5
250
312.5
375
437.5
500
Average Particle Diameter

A STATISTICAL EVALUATION OF THE PYROTECHNICS ELECTROSTATIC SENSITIVITY TESTER*

Everett Crane, Chester Smith, and Alonzo Bulfinch

OBJECT.

(a) To establish optimum operating conditions for the electrostatic sensitivity tester by determining statistically which design factors contribute most significantly to its performance.

(b) To determine whether electric spark sensitivity results obtained through use of this instrument on samples of fine (29 micron) magnesium powder are reproducible.

SUMMARY. An electrostatic sensitivity tester developed at Picatinny Arsenal was evaluated statistically. The factors found to contribute most significantly to optimum instrument operating conditions were resistance, humidity, energy, and the relationship of energy to resistance. The electrostatic sensitivity results obtained with fine magnesium powder specimens were found to be reproducible. It was concluded that further work should be conducted on a variety of samples to determine the effect of various characteristics of the circuit and the maximum energy input which will produce no burning in a specified number of trials. A method for measuring this can be developed by studying the lower tails of the spark sensitivity curves. Deviations in the lower tails of the curves, which are unique for each material, are the best indicators of the materials' sensitivity characteristics.

INTRODUCTION. Previously constructed electrostatic sensitivity testers were found to have one major shortcoming. The energy delivered to the sample was inconsistent because of losses within the system, and reproducibility of results was therefore erratic. An investigation of electrostatic sensitivity testers in use by the Bureau of Mines, the Naval Ordnance Laboratory, and the British armed forces was undertaken (Refs 1, 2, 3, and 4), and a modified apparatus was constructed in an attempt to eliminate this deficiency.

The action of the pyrotechnics electrostatic sensitivity tester developed at Picatinny (Fig 1)** is extremely simple. A sample is placed in the sample holder and a movable probe having a sharp point

---

* This paper appeared July 1959 as Dept. of the Army Project 504-01-027 report issued at Picatinny Arsenal.

** The figures have been placed at the end of this article.

is raised above it. The apparatus is then set at the desired voltage and R-C resistance. A chosen capacitor (charged to the desired voltage) is connected between the probe and the sample holder base. The capacitor is discharged by allowing the probe to fall to a fixed distance above the sample. The operator then observes and records the resulting action.

This is a classical experiment, as many such devices have been used in the past. However, despite its apparent simplicity, it has not, in past work, consistently produced satisfactory results. Because it has a built-in resistance, capacitance, and probe-down-time mechanism (Fig 2), the new device offers better opportunity for consistent results. One unfortunate difficulty, however, is that the probe (Fig 3) tends to become loosened by vibration, causing the operator to lose time in resetting it. After preliminary tests have been conducted, improvements to eliminate this fault will be made.

Because a large volume of data has been collected in determining optimum instrument operating conditions, it was considered desirable to issue a report on this phase of the investigation. Electric spark sensitivity data on various pyrotechnic, propellant, and explosive materials will be included in subsequent reports.

Difficulties inherent in the study of this instrument are:

1. Only attribute (Go, No go) type data can be obtained. This type of data yields only a small amount of information per observation.

2. The property of the materials to be tested is sensitivity to electric spark. This property requires a test of increased severity which is a type of test that yields little information per observation.

3. The effects of a large number of variables are determined simultaneously.

4. The spark sensitivity of a large number of materials must be evaluated. The input energy and the effect of instrument variables for any given material are of little value in the study of spark sensitivity of other materials.

5. Because of the nature of the data, non-parametric methods of analysis must be used. These methods are less efficient than parametric methods of analysis.

To reduce these difficulties to a minimum and extract the maximum amount of information possible, statistically designed experiments called factorial experiments were used. This type of designed experiment is the most efficient known. It is possible in factorial experiments to study more than one variable at a time. In general, the efficiency of the experiment is increased when a greater number of variables are studied simultaneously (Refs 5 and 8).

EXPERIMENTAL DESIGN AND ANALYSIS. Since the equipment used in this experiment was new, little was known at the outset concerning either the magnitude of the input energy required to cause burning or the effects of such other variables as might be present in the system. Therefore, a sequential approach to the problem was adopted. In this manner, something was learned about the magnitude of the input required, and it was possible to examine the results of small experiments before doing further work. The results of these exploratory experiments were not included in this report because their contribution was mainly to eliminate "rough spots" in the apparatus.

The data was analyzed by the Kruskal-Wallis rank-sum test, sometimes called the H-Test. In determining the significance of the main effects, this test was used in the usual way (Ref 5), to determine differences among means. In determining the significance of the first-order interactions, the appropriate main effects were subtracted from each total interaction effect.

In these exploratory studies, fine (29 micron average particle size) magnesium powder[1] was used, since it was a convenient homogeneous material.

Experiment 1 (Energy Changes)

To obtain a first estimate of the input energy required, tests of increased severity were conducted using the run-down method (Refs 6 and 7). In these tests, all variables were held constant at convenient levels, except energy (in joules), which was varied by varying the capacitance. When the results were plotted on probability paper (Figs 4 through 9), they yielded essentially straight lines, which indicated that the data could be considered, for all practical purposes, to be normally distributed. This was an important finding since it simplified interpretation of the results. The average values from these graphs (the 50% points in terms of energy) were helpful in establishing the input energy level used as a standard in subsequent experiments.

Experiment 2 (Gap Length, Humidity, Voltage, and Resistance)

The results of Experiment 1 were as follows:

1. The effects of sample size were insignificant.

2. Only inconclusive data was obtained on the effects of gap length and humidity.

3. The data obtained indicated that more should be known about the effects of voltage and resistance.

On the basis of the above findings, Experiment 2 was designed as a 4-factor complete factorial experiment to determine the effects of humidity, gap length, voltage, and resistance. The energy level was adjusted to 0.100 joule, to provide a usable distribution of successes and failures. The experiment was repeated 5 times (Tables 1, 2, and 3).

---

[1] Sample 142, barrel No. 30, Golwynne Chemical Company

Experiment 3 (Energy, Capacitance, and Voltage)

It was clear from the 4-factor experiment that the greatest number of ignitions were being obtained by eliminating the resistance (which is connected in series between the capacitor and the probe). It now appeared desirable to determine the effect of voltage at different energy levels. For this purpose, a 3-factor factorial experiment was designed (Tables 4 and 5) involving 3 levels of voltage, 6 levels of energy, and 2 levels of resistance. Resistance was included to confirm the conclusions reached in the 4-factor experiment regarding the effect of resistance.

RESULTS.

Experiment No. 1

The test s of increased severity showed averages (50% ignitions) and standard deviations (slopes), in joules, as follows:

	Average	Std Dev
Figure 5	0.100	0.075
Figure 7	0.134	0.055
Figure 9	0.144	0.064

Experiment No. 2

The results of the 4-factorial statistical analysis detailed in Tables 1, 2, and 3 were:

Main Effects[a]	Effect
Voltage (V)	Not Significant
Resistance (R)	Significant[b]
Gap Length (G)	Not Significant
Humidity (H)	Significant[b]

[a]Taken from the Analysis of Variance in Table 3

[b]Significant at the 95% confidence level

Interactions[c]	Effect
V x G	Not Significant
R x G	Not Significant
V x H	Not Significant

(contd.)

Interactions[c]	Effect
R x H	Not Significant
G x H	Not Significant
V x R	Significant

[c]Very highly significant, beyond the 99.9% level

Experiment No. 3

Figure 10 represents percentage of hits (burnings) versus volts versus joules and Figure 11 shows percentage of hits versus joules for 3000, 4000, and 5000 volts. The curve in Figure 12 is a composite of the 3 curves in Figure 11. Tables 4 and 5 show that, while resistance (R) and energy (E) are both very highly significant, voltage (V) is not significant. Figure 12 shows the average to be 0.062 joule and the standard deviation to be 0.019 joule over the three voltage levels used.

DISCUSSION OF RESULTS. Elimination of the danger of accidental electrostatic initiation is a major reason for measuring the electric spark sensitivity of pyrotechnics, explosives, propellants, and other materials. For this purpose, instrument operating conditions that will produce the maximum burning rate at all energy levels can be considered optimum.

From Tables 1 and 4, it is clear that removing all resistance from the system produces a significantly greater burning rate at all energy levels. Zero resistance can therefore be considered the optimum resistance condition for magnesium powder..

The data in Tables 4 and 5 and Figure 11 shows that, for zero resistance, the effect of changing the voltage from 3000 to 5000 volts is not significant. The effective sample size for evaluating the effect of voltage is 30 trials at each voltage level. Hence, the conclusion that the effect of voltage at zero resistance is insignificant at all energy levels is based on a sample size sufficient to give very good precision.

The data (Tables 4 and 5 and Figure 12) also makes evident a correlation between increasing percentages of burnings and increasing energy (joules).

Information on gap length and humidity is given in Table 1. This table shows that, over the 5 resistance levels, the effect of changing the gap length from 0.01 to 0.02 inch is nil and the effect of changing the humidity from 30% to 80% is significant. The results shown in this

table are considered to be reliable because they meet the effective sample size requirement for gap length and humidity, which is 250 trials at each level.

Additional work should be done to define the electric spark sensitivity of pyrotechnics, explosives, propellants, and other materials in terms of the characteristics of the electric circuit used and the maximum energy input which produces no burning in a specified number of trials. Once this definition has been developed through experience with representative materials, a method for measuring this property can be developed. This can be done by studying the lower tail of each sensitivity curve shown as a broken line in Figure 12. Since errors in this portion of the curve are rather large, it is dangerous to extrapolate from present data. In addition, significant deviations from normality can be expected. These deviations cannot be predicted by anv known means. However, past experience with the impact sensitivity of explosives has shown that these deviations in the lower tail of the sensitivity curve are unique for each material and are the best indicators of sensitivity characteristics.

Work should also be carried out to determine optimum instrument conditions for pyrotechnics, explosives, propellants, and other materials. It may be possible to classify most materials into a few general types for this purpose, so that only a few instrument settings will be required. If this is not possible, then a rapid method should be developed for determining optimum conditions for new materials.

CONCLUSIONS.

1. The maximum burning rate of magnesium powder cannot be obtained over the range of energy levels surveyed if resistance is added in series between the capacitor and the probe. Varying the voltage between 3000 and 5000 volts has no effect on the number of ignitions of magnesium powder at any energy level when the resistance level is held constant.

2. Ignition is dependent on the energy released by the electrostatic sensitivity apparatus. For magnesium powder, the percentage of burnings increases with increasing energy (joules).

3. There is highly significant interaction between resistance and voltage, that is, the effect of voltage is dependent upon the level of resistance employed. Thus, any statement concerning the effect of voltage on burnings must specify the level of resistance.

4. The electrostatic sensitivity results obtained for 29-micron-average-particle-size magnesium powder are reproducible.

5. Additional work will be needed to evaluate the effect of gap length and humidity at zero resistance and to determine the electric spark sensitivity of a wide range of pyrotechnics, explosives, and propellants.

TABLE 1

Four-Factor Factorial Electrostatic Sensitivity Experiment[a] (Experiment 2) for 29-Micron Magnesium Powder

Relative Humidity	Gap Length, inches	Resistance, kilo ohms[b]	E_1 .0222 mfd[c] 3000 volts[d]	E_2 .0163 mfd 3500 volts	E_3 .0125 mfd 4000 volts	E_4 .0099 mfd 4500 volts	E_5 .0080 mfd 5000 volts	Total Hits
25–40%	0.021	0	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	0 1 1 1 1	1 1 1 1 1	24
"	"	90	1 0 0 1 1	1 0 1 1 0	1 1 0 0 0	0 1 0 1 1	1 0 1 1 0	14
"	"	170	1 1 0 0 1	0 1 0 1 1	1 1 1 0 0	1 0 0 1 0	1 0 1 0 1	14
"	"	260	0 1 0 1 1	1 0 1 0 1	1 1 1 0 0	1 0 1 1 0	0 1 1 1 0	15
"	"	350	1 1 0 1 1	0 0 1 1 1	1 0 0 1 1	1 0 1 0 1	0 1 1 0 0	15
"	.010	0	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	25
"	"	90	1 1 1 0 1	1 1 0 0 1	0 0 0 1 1	0 0 1 0 1	1 0 0 1 0	13
"	"	170	0 1 1 0 0	0 0 1 0 1	1 0 0 1 0	0 1 1 1 0	0 1 1 1 0	12
"	"	260	1 1 0 1 0	0 1 1 1 0	1 0 1 0 1	0 1 1 1 1	1 1 1 0 1	17
"	"	350	1 0 0 1 0	0 1 0 0 1	0 0 1 0 1	0 1 1 1 1	0 0 1 1 1	13
75–95	.021	0	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	25
"	"	90	1 1 1 1 0	1 0 1 1 1	0 0 1 1 0	1 0 1 1 0	1 1 0 0 1	16
"	"	170	1 1 0 1 1	0 1 1 1 1	1 0 1 1 0	1 1 1 1 0	0 1 0 1 1	18
"	"	260	1 1 1 1 0	0 1 1 0 1	1 0 0 1 1	1 1 0 1 0	1 0 1 1 1	17
"	"	350	0 1 1 1 1	1 1 0 1 0	1 0 1 1 0	1 1 1 0 1	1 1 0 1 0	17
"	.010	0	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	1 1 1 1 1	25
"	"	90	1 0 1 1 1	1 1 1 0 1	1 0 0 1 1	0 1 0 1 0	1 1 1 1 0	17
"	"	170	1 1 1 0 1	1 1 1 1 1	1 0 0 0 1	1 0 1 0 1	1 1 0 1 0	17
"	"	260	1 0 1 1 0	1 0 1 0 1	1 1 0 1 0	1 0 0 0 1	0 0 1 1 1	14
"	"	350	1 1 0 1 1	1 1 1 0 0	1 1 1 0 1	1 1 1 1 1	1 1 1 0 0	19

[a] Energy, $E = \frac{1}{2} C V^2 = 0.100$ joule at every level; probe dwell-time 2.5 seconds; 2 standard scoop quantities. 0 = No Reaction; 1 = Reaction.

[b] From R–C resistance (See Table 2, p 8).

[c] Capacity

[d] Voltage

TABLE 2

Summary of Table 1 Data

	Trials	Hits	Misses
Capacitance and Voltage*			
E_1	100	75	25
E_2	100	71	29
E_3	100	63	37
E_4	100	69	31
E_5	100	69	31
Resistance, ohms			
0	100	99	1
90,000	100	60	40
170,000	100	61	39
200,000	100	63	37
350,000	100	64	36
Gap Length, inches			
.021	250	175	75
.010	250	172	78
Humidity, %			
25 to 40	250	162	88
75 to 95	250	185	65

	E_1	E_2	E_3	E_4	E_5
*Capacitance, mfd	.0222	.0163	.0125	.0099	.0080
Voltage	3000	3500	4000	4500	5000

Energy was in all cases .100 joule.

TABLE 3

Non-Parametric Analysis of Variance of Table 1 Data

	Calculated H-value[a]	Degrees of Freedom	Critical Chi-Square
MAIN EFFECTS			
Voltage (V)	3.3	4	9.49
Resistance (R)	11.7[b]	4	9.49
Gap Length (G)	0.0	1	3.84
Humidity (H)	4.8[b]	1	3.84
INTERACTIONS			
V × G	2.3	9	16.92
R × G	12.5	9	16.92
V × H	0.0	9	16.92
R × H	14.5	9	16.92
G × H	2.0	3	7.81
V × R	85.9[c]	24	36.42

[a] $H = \frac{12}{N(N+1)} \sum_{1}^{k} \frac{(R_i)^2}{n_i} - 3(N+1)$. This H-test is the Kruskal-Wallis rank-sum non-parametric test for the difference among means of counted data where H has a Chi-square distribution and

N = Total number of determinations in all groups ($\Sigma n_i = N$)

k = Number of groups

n_i = Number of determinations in an individual group

R_i = Sum of the ranks in an individual group.

[b] Significant at the 98% level

[c] Very highly significant

TABLE 4

Three-Factor Factorial Electrostatic Sensitivity Experiment[a,b] for 29-Micron Magnesium Powder

Energy, joules	Capacitance, microfarads	Voltage, kilovolts	Trials[c]	Total Hits
0.10	0.0222	3	1 1 1 1 1 1 1 1 1 1	10
"	.0125	4	1 1 1 1 1 1 1 1 1 1	10
"	.0080	5	1 1 1 1 1 1 1 0 1 1	9
.08	.0178	3	1 1 1 1 1 1 1 0 1 1	9
"	.0100	4	1 0 1 1 1 0 1 1 1 1	8
"	.0064	5	0 1 0 1 1 1 1 1 1 1	8
.07	.0155	3	1 1 1 1 0 1 1 0 1 1	8
"	.0088	4	0 1 1 1 1 0 0 1 1 1	7
"	.0056	5	1 1 1 1 0 1 0 1 0 1	7
.06	.0133	3	1 1 0 1 0 1 0 0 1 0	5
"	.0075	4	0 1 0 0 0 1 0 1 1 1	5
"	.0048	5	0 1 0 1 0 0 1 0 1 0	4
.05	.0111	3	0 0 1 0 0 1 0 1 1 0	4
"	.0063	4	0 0 1 0 0 1 0 0 0 1	3
"	.0040	5	0 0 0 0 0 0 1 0 0 0	1
.04	.0089	3	0 0 0 1 0 1 0 0 0 1	3
"	.0050	4	0 0 0 0 0 0 0 0 0 0	0
"	.0032	5	0 0 0 0 0 0 0 0 0 0	0

[a] This experiment was repeated for 10,000 ohms R–C resistance with 100% failures (No reactions). See Table 5 (p 11).

[b] Probe dwell-time 2.5 secs, R–C resistance = 0 ohms, Gap length 0.01 to .02 in.; R. H. 25 – 35%

[c] 0 = No reaction; 1 = Reaction

TABLE 5

Summary of Table 4 Data (See also Figs 8 and 9)

Energy, joules	Voltage	% Hits	
		Zero Resistance	10,000 ohms Resistance
.10	3000	100	20
.10	4000	100	0
.10	5000	90	0
.08	3000	90	0
.08	4000	80	0
.08	5000	80	0
.07	3000	80	0
.07	4000	70	0
.07	5000	70	0
.06	3000	50	0
.06	4000	50	0
.06	5000	40	0
.05	3000	40	0
.05	4000	30	0
.05	5000	10	0
.04	3000	30	0
.04	4000	0	0
.04	5000	0	0

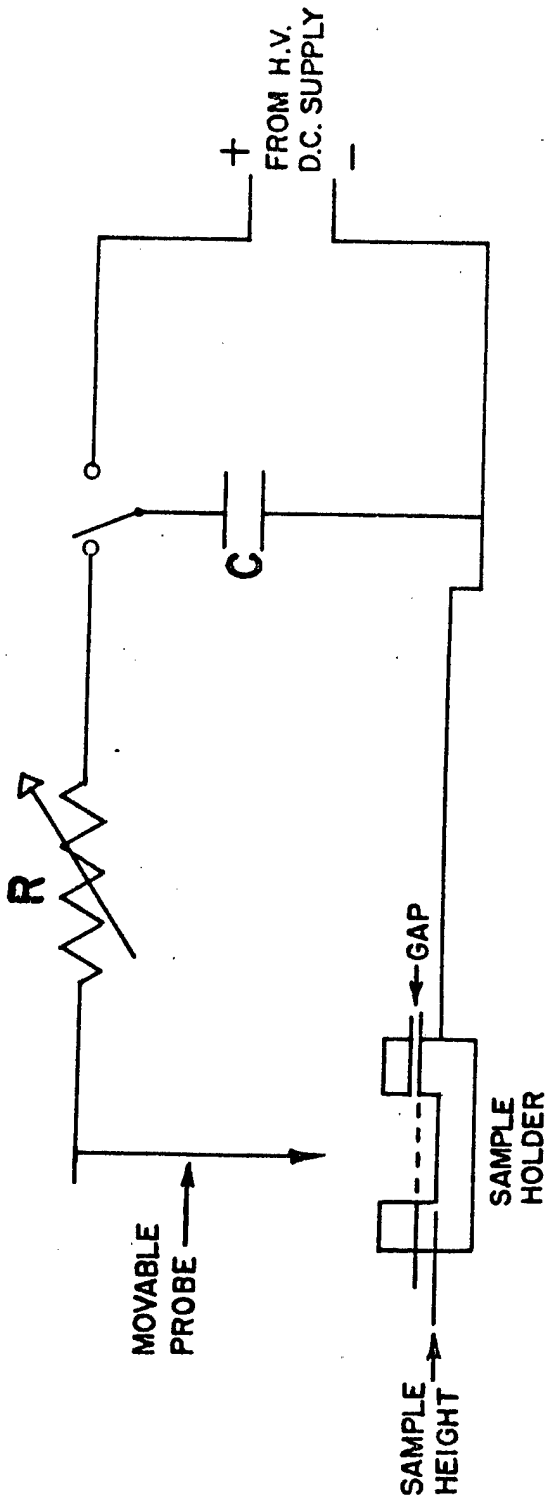


Fig 1 Schematic of Electrostatic Sensitivity Test Apparatus

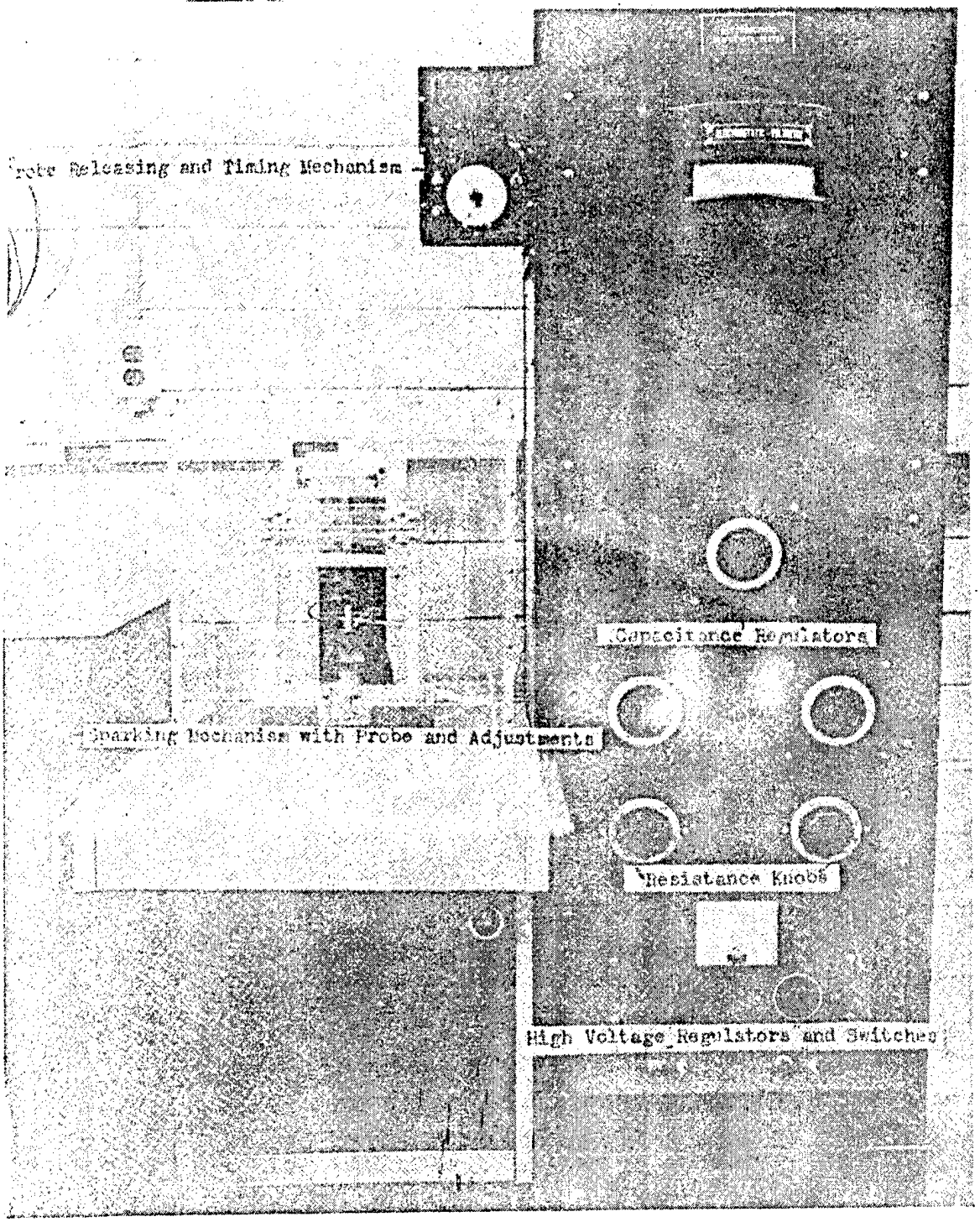


Fig 2 Pyrotechnic Electrostatic Sensitivity Tester

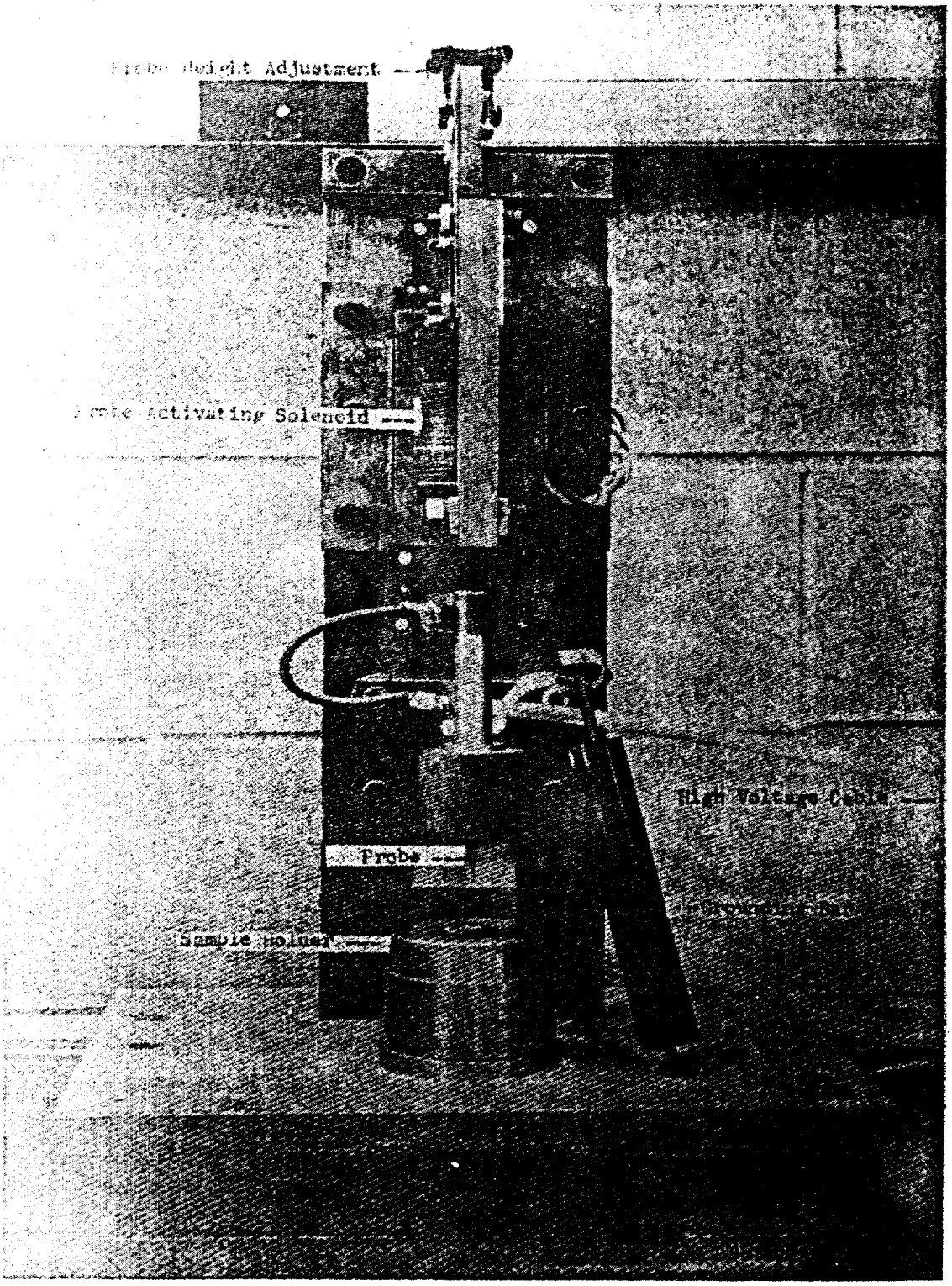


Fig 3 Sparking Mechanism, Probe, and Adjustments

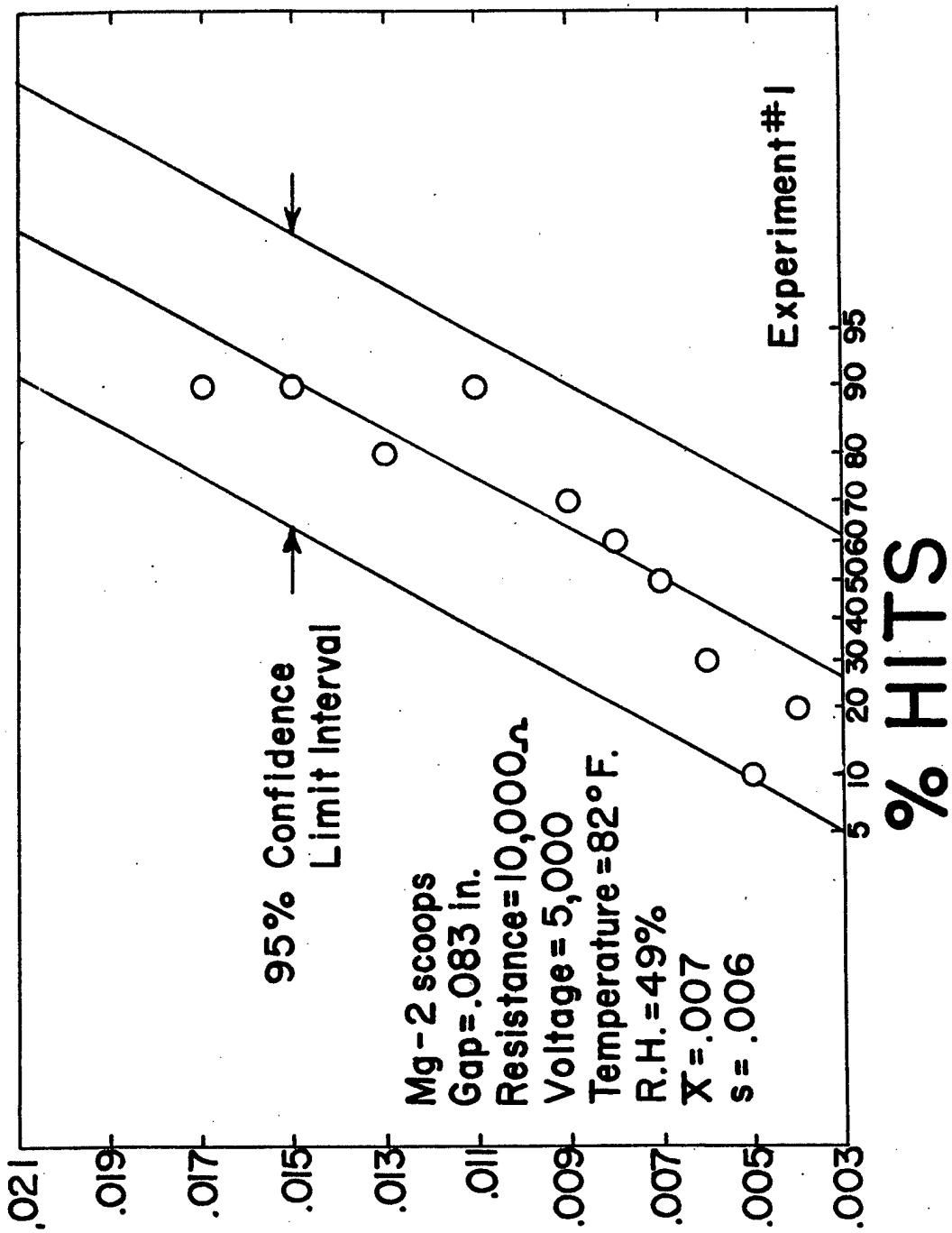


Fig 4 Normality of Distribution of Experiment 1 Capacitance Data

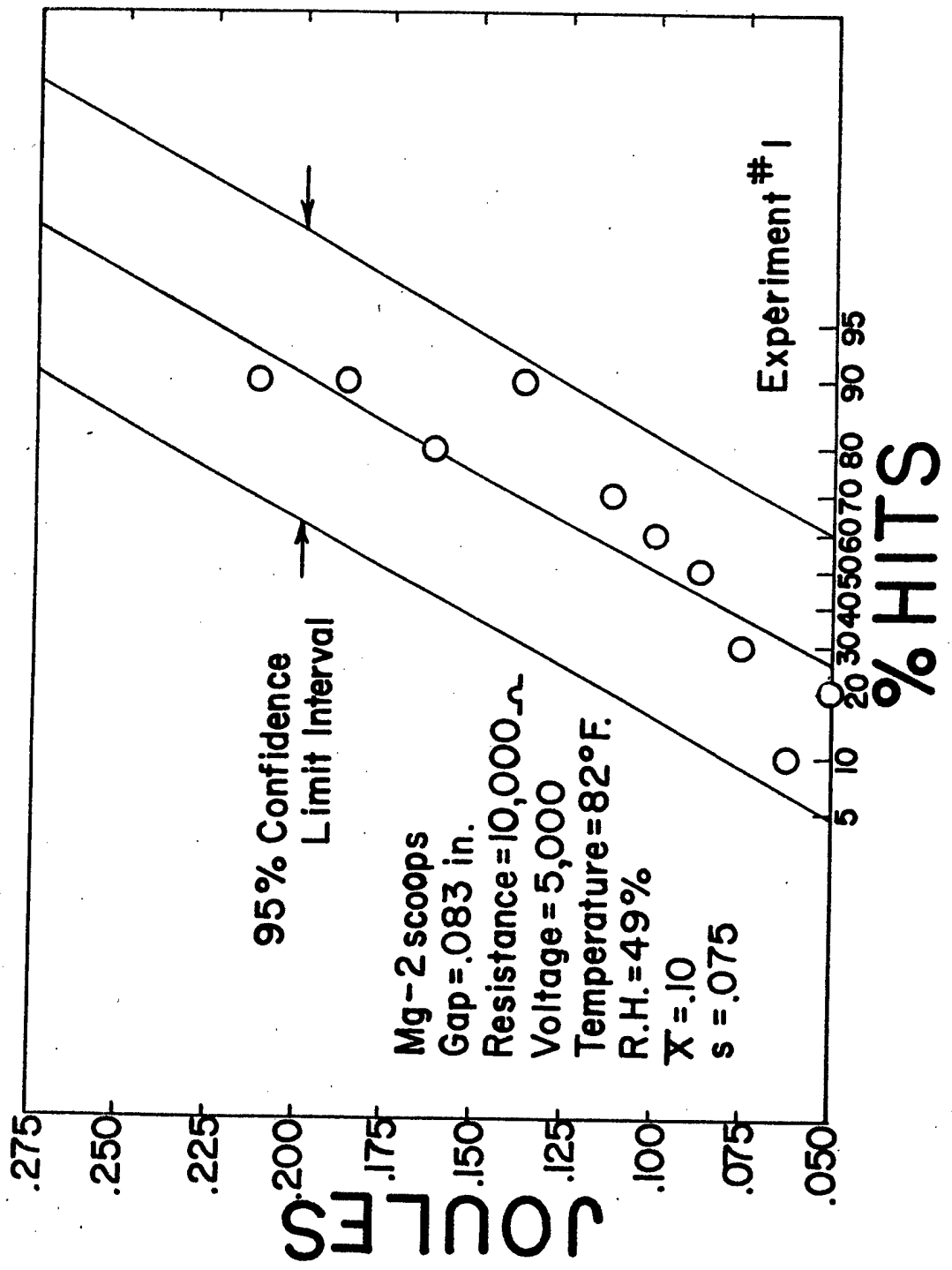


Fig 5 Normality of Distribution of Experiment 1 Energy Data

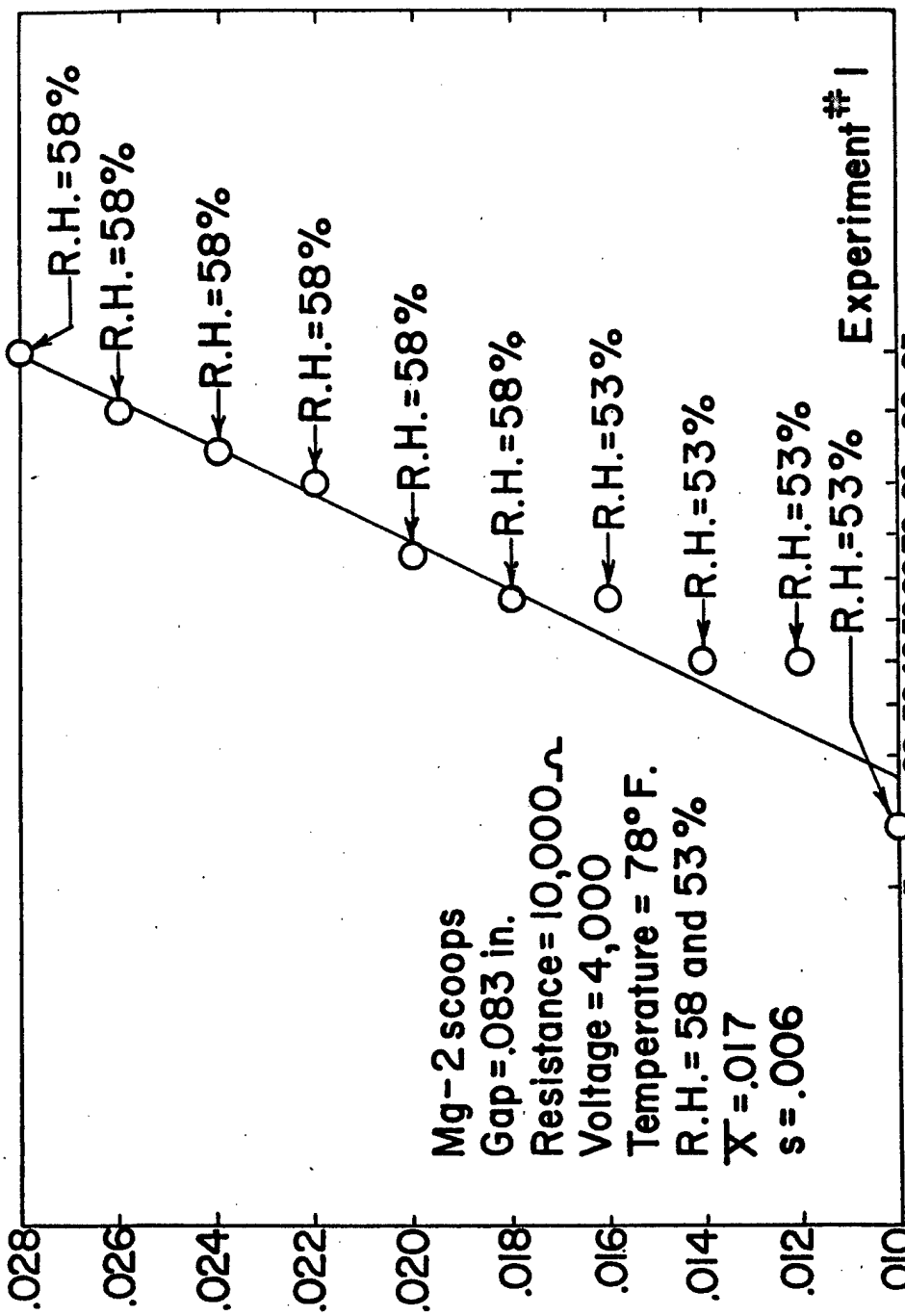


Fig 6 Preliminary Data on Effect of Humidity on Capacitance

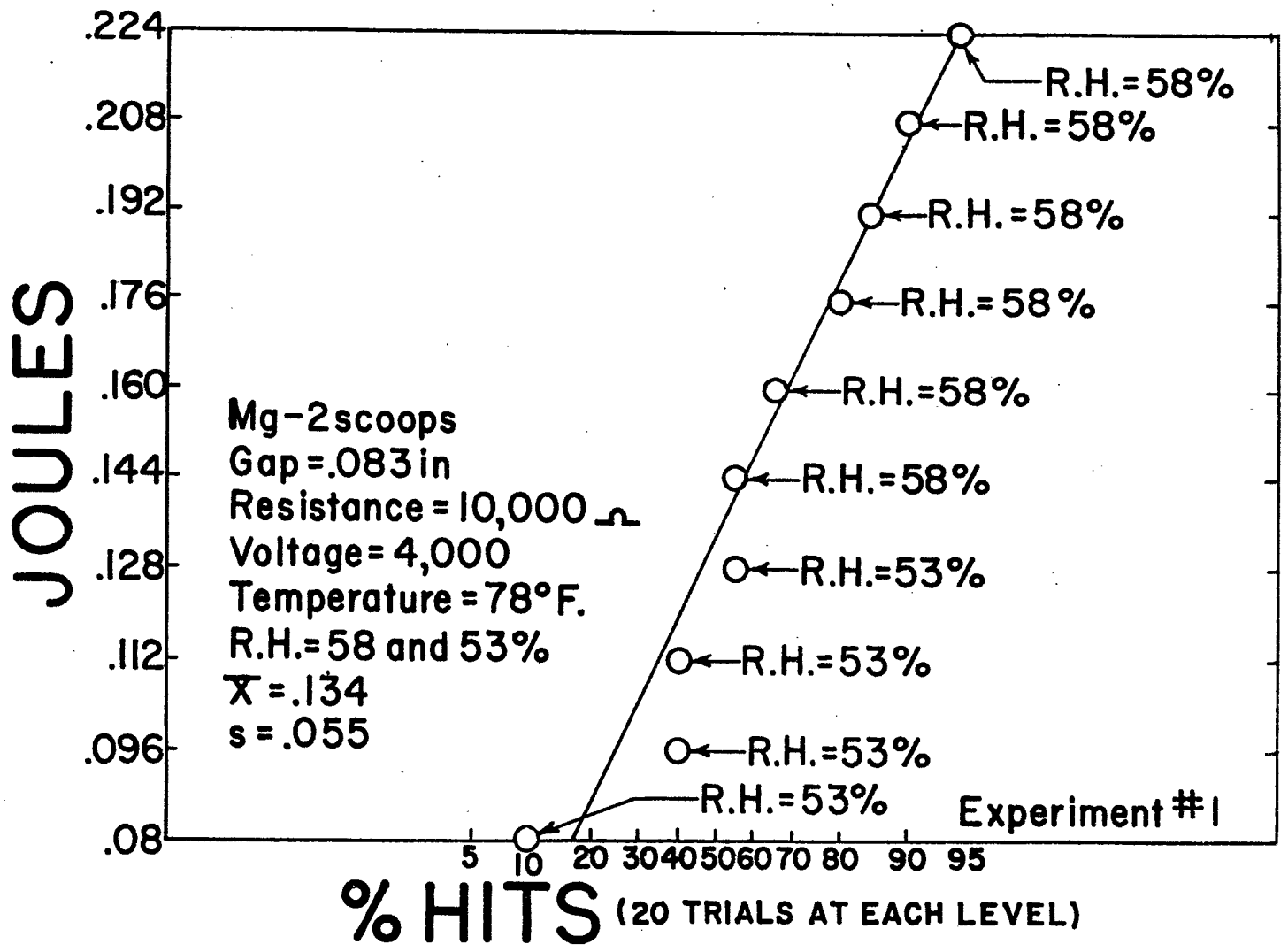


Fig 7 Preliminary Data on Effect of Humidity on Energy

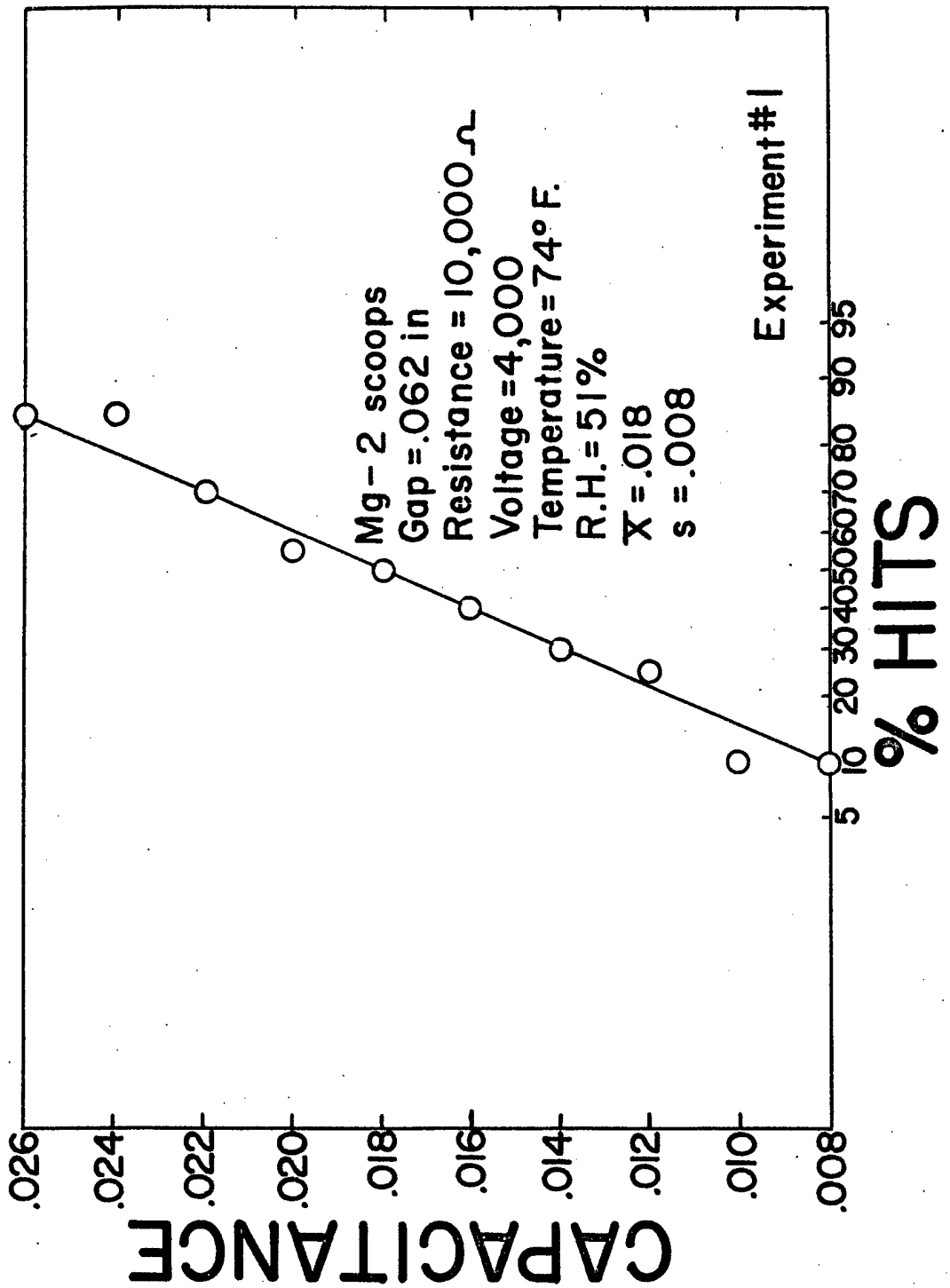


Fig 8 Confirmation of Figures 4 and 6 at Controlled Temperature and Humidity

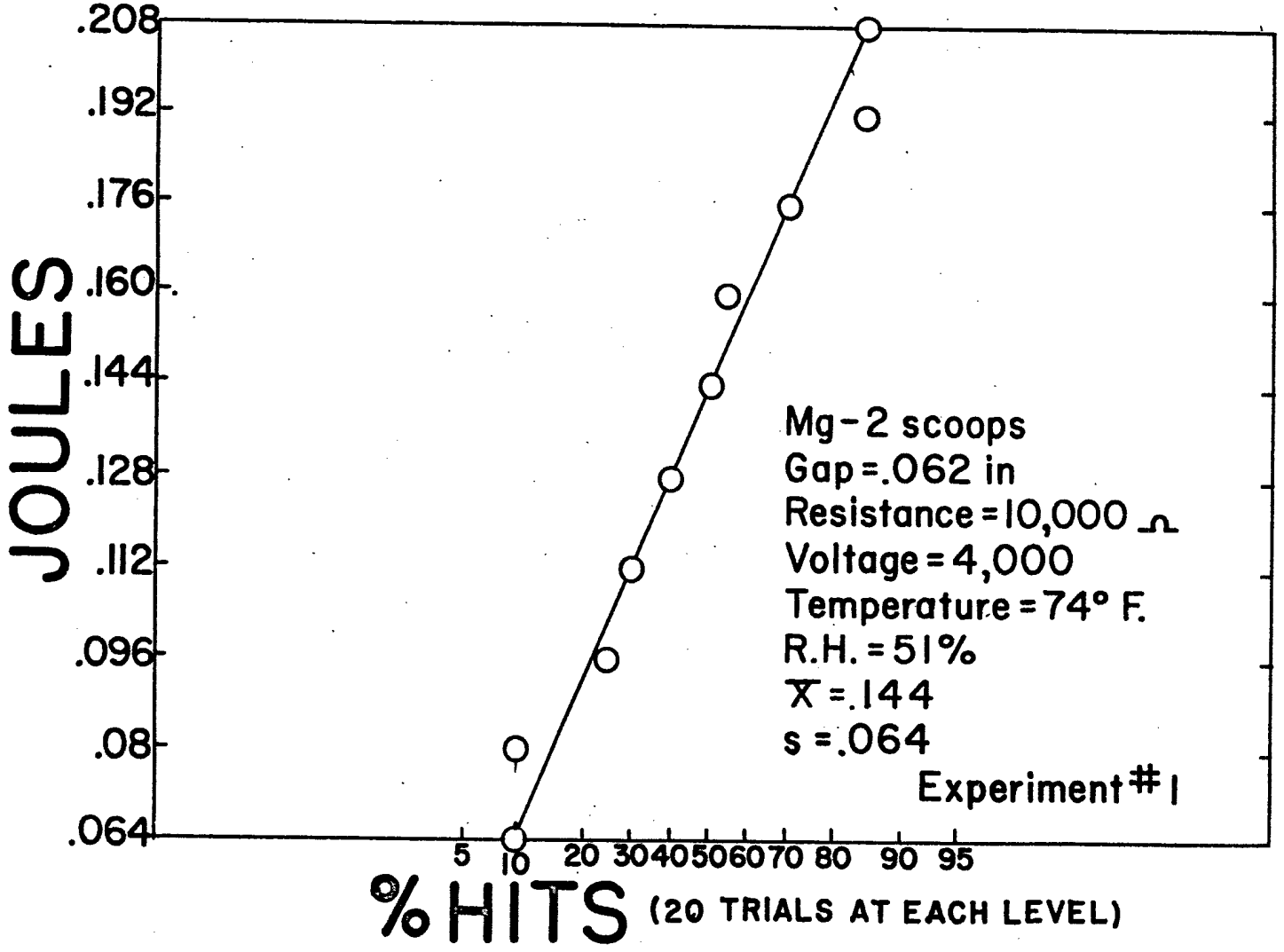


Fig 9 Confirmation of Figures 5 and 7 at Controlled Temperature and Humidity

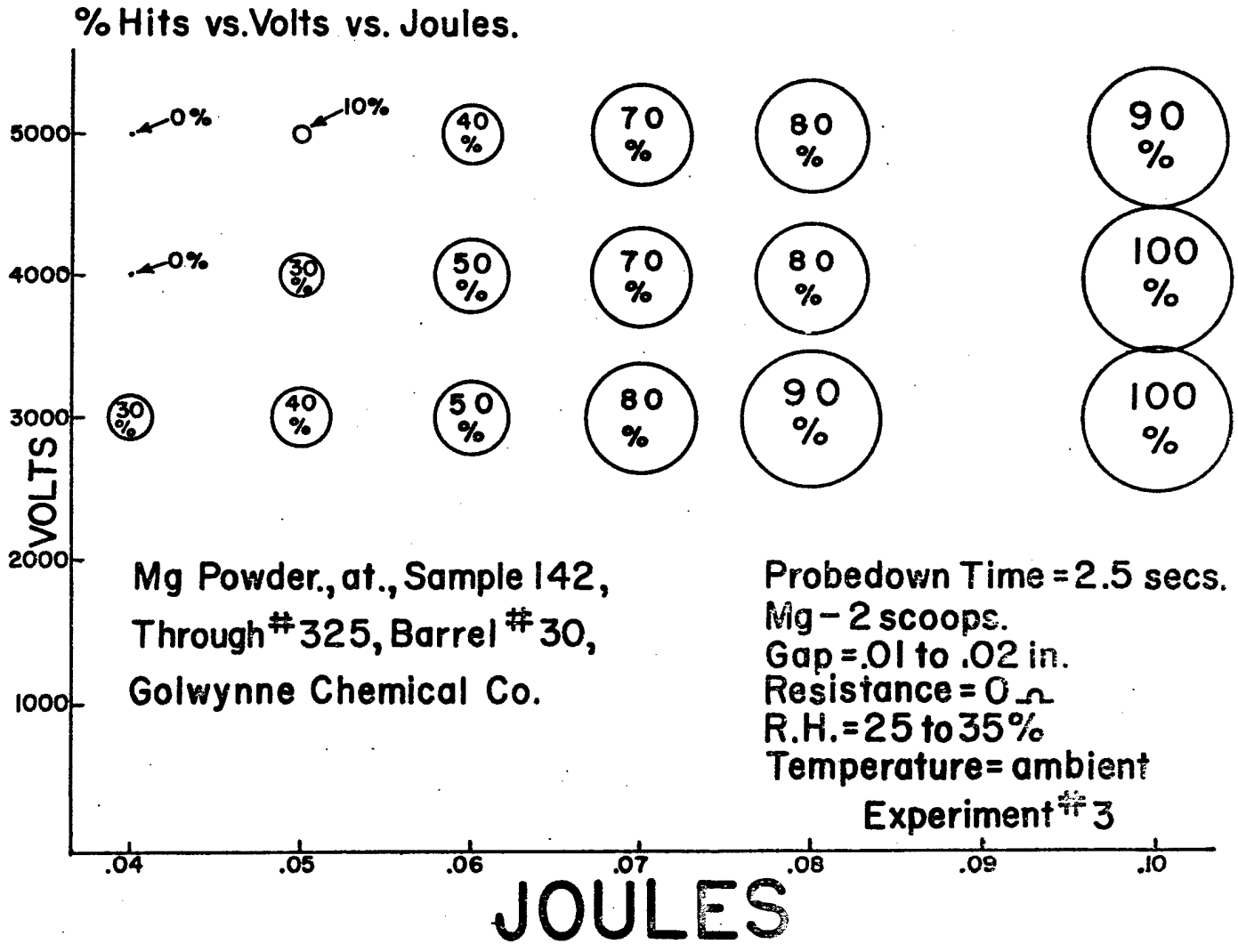


Fig 10 Area Graph Showing Interaction Between Voltage and Energy

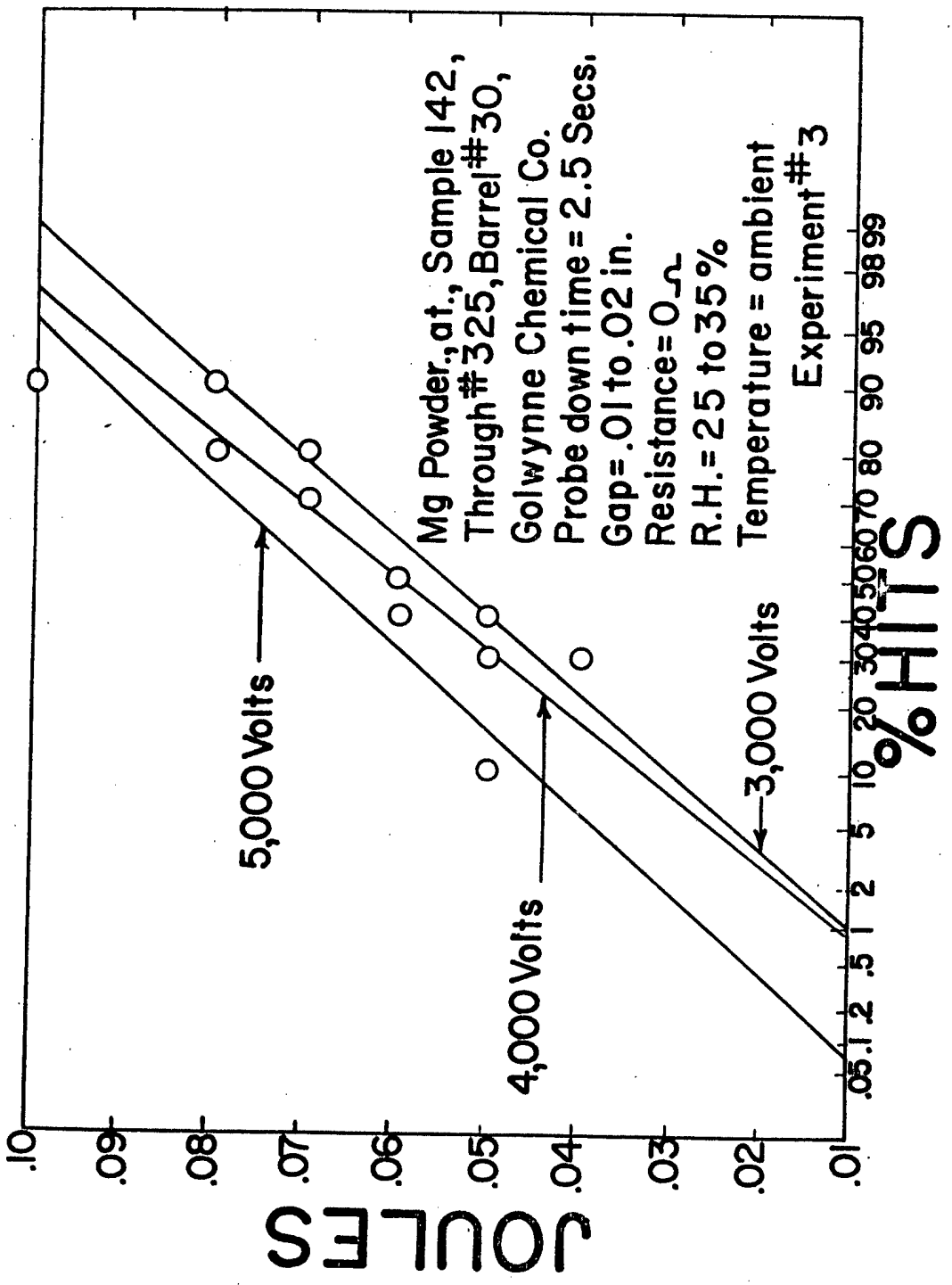


Fig 11 Line Graph Showing Interaction Between Voltage and Energy

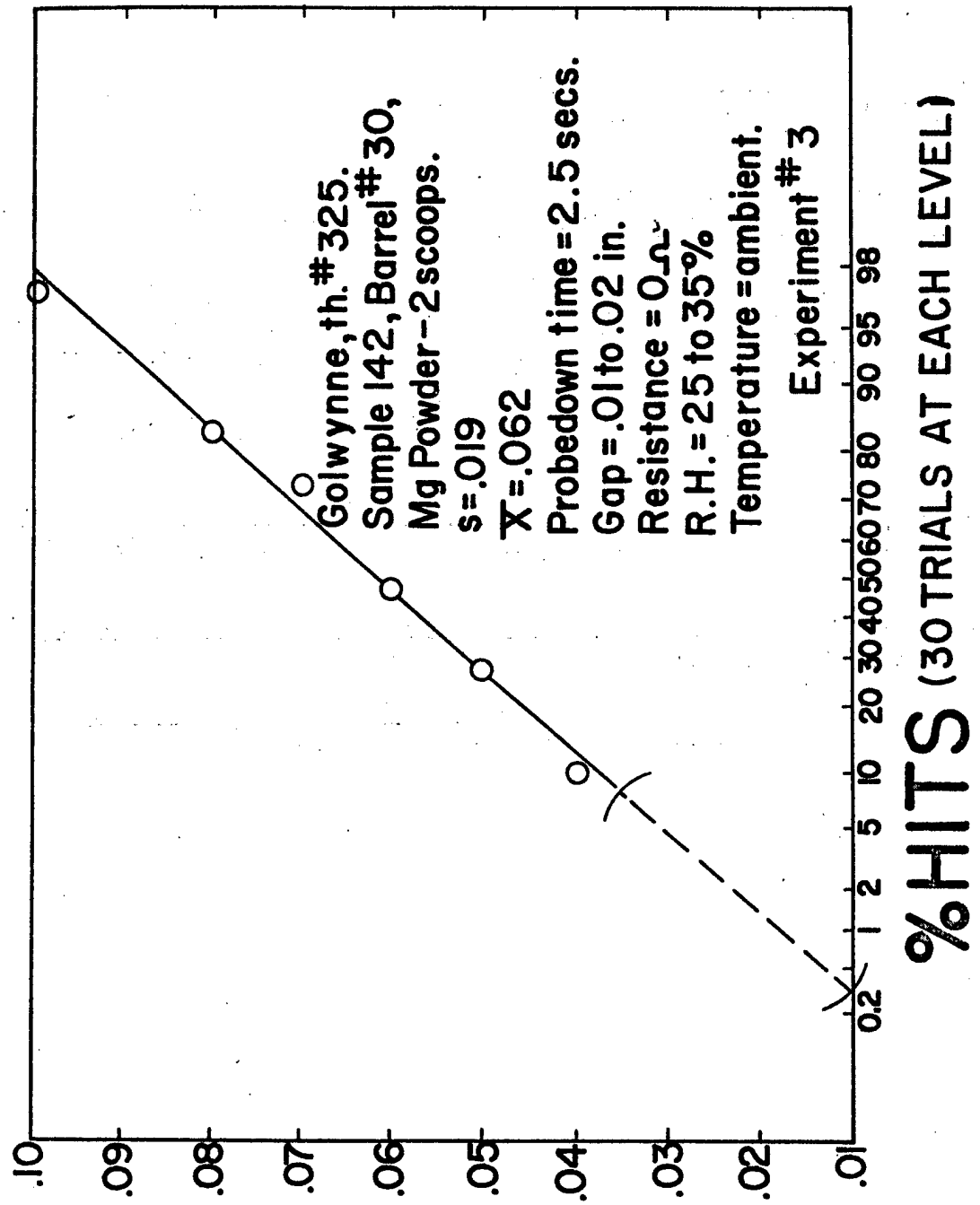


Fig 12 Composite of Figure 11 Curves

REFERENCES

1. I. Hartmann, J. Nagy, and H. R. Brown, Inflammability and Explosibility of Metal Powders, Bureau of Mines, R. I., 3722, October 1943

2. J. N. Ayres, The Design, Assembly, and Operation of the Explosive Electrostatic Sensitivity Tester, Naval Ordnance Laboratory Memo 9959, 7 Feb 1949

3. F. W. Brown, D. J. Kusler, and F. C. Gibson, Sensitivity of Explosives to Initiation by Electrostatic Discharges, Bureau of Mines Report 5002, September 1953

4. P. W. J. Moore, J. F. Sumner, and R. M. H. Wyatt, The Electrostatic Spark Sensitiveness of Initiators: Part 2 - Ignition by Contact and Gaseous Electrical Discharges, C35838(10), May 1956

5. W. Dixon and F. Massey, Introduction to Statistical Analysis, 2nd Edition, McGraw Hill Book Co., Inc., New York City, 1957, page 290

6. C. W. Churchman, Theory and Application of Sensitivity Curves of Small Arms Primers, as Determined by the Standard Drop Test Machine, Frankford Arsenal Report R-259, December 1942

7. C. W. Churchman, Manual for Proposed Acceptance Test for Sensitivity of Percussion Primers, Frankford Arsenal Report R-259A, January 1943

8. O. L. Davies, The Design and Analysis of Industrial Experiments, Hafner Publishing Co., New York City, 1954

ABSTRACT

EXPERIMENTAL DETERMINATION OF "BEST"
COMPONENT LEVELS IN THERMAL POWER SUPPLIES

Sheldon G. Levin
Diamond Ordnance Fuze Laboratories

The paper presented the results of a study conducted by the Power Supply Branch of the Diamond Ordnance Fuze Laboratories. The primary purpose of the experiment was to find the combination of component weights (levels) which would give optimum performance of a particular battery under any use (test) condition. Three factors A, B, C, were considered at four levels each, and the details leading up to the selection of a 3/4 replicate were given. There were four responses: life, activation time, noise level, and peak volts, each examined at four test conditions. It was necessary to establish criteria of goodness and state the objectives in mathematical form.

There was no exact solution to this multivariate problem and the fact that the two different approaches to the analysis gave very similar results was very encouraging. The first explored each of the response surfaces and selected the "best" region by ordering the standardized form of the responses. The second method obtained a linear function of the factors which gave a single continuous variable $f(x,u,z) = f$. The responses of life, activation time, peak volts and noise level were then fitted as quadratic functions of f. The set of values of f which corresponded to the satisfactory (with regard to end-use requirements) values of each response variate was obtained. The intersection of these sets constituted the satisfactory, and in this case best region.

STATISTICS IN MEDICAL RESEARCH

W. J. Dixon
University of California, Los Angeles

I shall not attempt an introduction to the statistical methods used in the medical field. I shall attempt to discuss some of the special conditions which the nature of the field imposes on the design situation. As in any experiment we need to know something about the observer. I started out in mathematics, specialized in mathematical statistics, then worked with applications in engineering and social science, and now work in the biomedical and health sciences.

What is the professional environment of medical research? Who are the people conducting this research? They are:

1. Ph.D.'s from physical sciences working in biophysics in instrumentation and simulation research

2. Ph.D.'s from chemistry and physiology doing experimentation in pharmacology, psychiatry and chemotherapy.

3. M.D.'s of all specialties carrying out animal and plant work as well as clinical trials on humans

What problems are investigated? Problems studied in the medical research environment range from the most basic kinds of research of the type carried out in physics, chemistry and physiology and as carried out in psychology and sociology to mere collections of observations--perhaps, somewhat objective--on standard treatments in standard medical practice. Some form of what we call "statistics" is present in all of these types of research.

What are examples of differences in medical applications? First, consider survey research. Surveys involving records only may present only the usual difficulties in probability sampling, stratification, etc.* But a serious difference may arise if direct interviews or questionnaires concerning health or previous medical history are undertaken. Here, additional problems of non-response and language aries (the National Health Survey is developing techniques in this area). If we seek mental health status, we find problems which at present seem insurmountable. (Nothing but plans that I know of). However, some of the most important differences and differences which force some adjustment arise from the fact that people are involved.

How are people involved?

1. Directly as experimental subjects

---

* There is an excellent article discussing these applications in Am. J. Ph H. 44 (1954) pp. 719-740, On the Use of Sampling in the Field of Public Health.

Or we may be involved

2. Only slightly less directly if the research outcome has an immediate effect on our own disease or on the diseases of those close to us or on a disease we might expect to contract.

3. We feel involved as a member of the human community seeking a cure or being concerned about experimentation on humans.

4. Even if there were no personal concern about ultimate results there are laws governing the treatment of human beings. Roughly stated the law defines the correct treatment as the treatment in general use.

1. Each individual is concerned about his α and β risks (even though they may be poorly formed in his mind) when he thinks about being included in a study. How can one obtain a random sample of objects when they can choose to be unsampled?

2. If an individual's own life is to be greatly affected in a different way depending on what a study shows, how can he be objective?

How can one do careful and well planned research when so much pressure for time is brought to bear on laboratories connected with medical research? Even though work is still at the stage where it is being carried out on animals, or biological systems other then man, this research is often considered only an early step to later research of direct relevance to people--thus the pressure is on at all stages. Conclusions must be obtained quickly before assumed knowledge is prevalent--there may be no possibility for slow efficient sequential experimentation, particularly on man, because in the medical profession information moves rapidly from mouth to mouth, and if a treatment becomes generally accepted, whether rightly or wrongly, research cannot continue. In the urge for speed we may find the use of a control resisted due to the feeling: "If the drug works we can try it on twice as many people during the same period." As an exmple, radiation is regularly used as a supporting treatment to surgery for certain cancers. This is not experimental, but accepted. It was not given large scale scientifically designed trials before becoming accepted. This would now be very difficult to do. Since chemotherapy in addition to surgery was not accepted previously, experimentation is possible. Such research is in progress.

Another point which needs careful statement concerns the clinical result versus scientific or statistical result. At least at first, the clinician shows little interest in separation of two groups in <u>mean</u> even though the separation may be precise and real; in clinical work the immediate concern is with each single person. The question is usually asked: Is this result of clinical significance? This may mean can you classify <u>individuals</u> into one of several categories with minimum error?

In a very similar way difficulties arise in the choice and definition of measurement to be used in an investigation. In the basic sciences there is early attention to refinement of criteria, measurement such as refined weighing scales, an accurate radiation counter or some accurate electrical measurement. In the behavioral sciences there is often extensive development of test or questionnaire or interview procedure to gain adequate reproducibility of measurement. In medical research the same care must be taken. You may find that a technique for measuring blood pressure as carried out in office practice (which is entirely adequate for deciding whether a patient's blood pressure is closer to 250 than it is to 120) may immediately be used without refinement in a research study to investigate changes over short periods of time caused by small doses of tranquilizers. Of course, the measurement will almost surely fail to detect these differences. There is a common rationalization supporting the use of unrefined measures in medicine which arises from the confusion of the discovery of a new result and the later developmental problem of making a result of practical importance. The clinician may state: "If the result is not observable with the usual techniques, it will not be of practical value."

In addition to errors in the measuring instrument itself, which are often large, it may be necessary to refine the measurement to some basal state for the individual. One may be able to design a study with sufficient replication of measurement on enough individuals so that one need not hold fixed some of the greatest contributions to variation, but this is usually not the case. In the case of blood pressure for the basal state the recommended technique requires twelve hours of the patient's time and two or three hours of the clinician's time. On the other hand, rapid reading may have tremendous variation. Research is needed in developing for many measurements compromises which will result in more accuracy without too much cost. For the example of blood pressure, some workers are investigating the use of a reading taken a short time after giving a tranquilizer. Development in this direction, of course, introduces a new problem. For example, ths use of a tranquilizer may change blood pressure differentially for the type of patient you wish to discriminate so that he is not separated from others.

It must be generally recognized that the type of measurement used should be chosen on the basis of the research goals. However, some research workers with a clinical background may understand that research will require different measurements than are used in clinical practice, but may seek more and more accuracy, when they may need only precision. In some cases precision may be available by a presently known or easily developed technique. If they are investigating changes only, precision may suffice. For example, they may need to observe only an upward or downward shift and be little concerned with absolute level.

Present Medical Record System Creates Problems for Research: The collection and storage of certain measurements and observations on patients is required for various legal and accreditation requirements. In many places the clinician knows that these records are never used in more than a superficial way. Even though certain data must be coded

accurately, much data need not be. Much of his experience in recording data is with these medical records. Now, suppose you organize a study and require from a certain physician measurements on a patient. His first inclination will be to supply information in the form which appears in the medical records. This information may be of very little value for the research study either because of the use of very gross categories or because the desired information may be included for only certain types of cases. In some cases it can be important to note the presence or absence of the information rather than the size of the measurement if actually recorded. There are many record systems in hospitals and clinics. Records are kept by admissions, departments, by nurses, operating room staff, clinical laboratories, individual physicians. They are kept for medication, infections, special procedures, etc. When a research problem is instituted and a new form is introduced it may receive the easy cavalier verbal fluency of some present records rather than the persevering scrupulous accuracy necessary for good research.

Difficulty of Measuring One Component or Holding Other Variables Fixed: Much of medical research is done on living organisms of a complex nature, man, for example. Any measurement may affect the individual so that immediate replication is often impossible. The body has many compensating systems so that measurement of only one characteristic is of little value. The component cannot usually be measured except when coupled to all other components of the body. Selective assembly of components is usually not possible. Progress may be made for large or specific types of response with few measurements, but frequently the only approach is through a multivariate analysis.

To what question does the medical research worker seek answers? He asks what to measure. The answers should not only give consideration to the accuracy of each particular measurement but the choice of which to use. It is recognized that many measurements will be required, but one cannot measure everything for reasons of time and and money alone. Therefore, studies or redundancy of observations may be required. A component analysis of basic measurements may be helpful. A component analysis may also be helpful in the reduction of the number of variables which must be considered in solving the diagnosis problem.* Regression analyses on basic variates or on components may allow a reduction in the number of variates of importance for certain types of research problem. In other words, we may be able to help in determining which independent variables are of importance for different dependent variables.

There is the question: How to measure? For measurements on a continuous scale the effectiveness of a particular measurement in the analysis of an experiment may be increased by the use of a

---

* Discussions of these techniques are given by:
M. G. Kendall, A Course in Multivariate Analysis, Hafner, New York, 1957
S. S. Wilks, Mathematical Statistics, Princeton, N. J., 1943

transformation.* Many observations made in medical research are ordered but without a natural scale. For example, severity of response, answers to history questions, symptoms, laboratory findings, response to treatment. The effective use of such variables will usually require the assignment of a numerical scale. Decision will be required as to criteria underlying the scale to be chosen. One may choose a scale which will optimize a certain regression relationship or one may choose a scale to minimize interaction with another observation being made in the analysis.** The literature dealing with problems of scaling is certainly not complete, but see Torgerson.***

In contrast with this discussion consider clinical trials. Here one may attempt to control many variables not by multivariate analysis but by randomization. For example, one can accept cases as they arise and randonly assign them to treatment or control categories and trust the randomization to effect a balance on the many other related conditions. If knowledge of the status can be kept from the patient and from the doctor these studies are called "double blind." Since the treatment often involves a specific act by the doctor a placebo or dummy treatment is often used.
Examples: Heart surgery, chemotherapy, cold treatment, etc.

What does medical research need (from statisticians):

1. Improvements in basic measurements including scaling methods.

2. Computer programs to ease the pain of multivariate analysis.

3. Further developments in analysis of multivariate measurements including those made continuously in time, e.g., spectral analysis.

How to Use Statistical Methods in Medical Research?

As can easily be seen from reading many current periodicals in the medical field, a great many papers use statistical findings. One can also note that the words probability, confidence and significance are used, but what has been attempted by the statistical analysis often could be called resurrection or sanctification. Since there is a tendency for statistical aid to be requested for the poorer study at the wrong time, it is important to show good research workers how modern statistical methods can be a part of their entire "grand strategy" of the use of the scientific method.

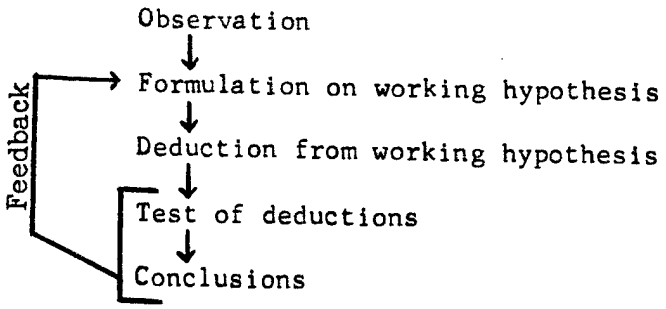
---

*For example, see John W. Tukey, On the Comparative Anatomy of Transformations, Annals of Math. Stat., Vol. 28 (1957), p. 602

**An interesting application of this criterion for scaling severity of response appeared in a paper by P. J. Claringbold and W. R. Sobey, Studies on Anaphylaxis, Australian J. of Biological Sciences, 10 (1957), 360-364.

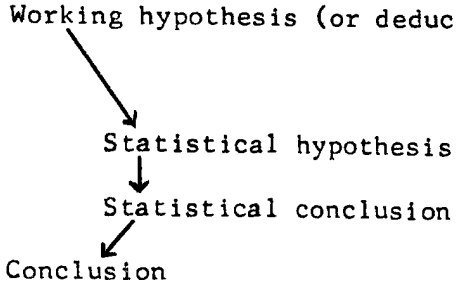
***Warren S. Torgerson, Theory and Methods of Scaling, John Wiley & Sons, New York, 1958.

If we think of the outline



we can perhaps show the research worker that the use of statisitcal methods can make more efficient at least one stage in the above diagram enabling the experimenter to more to the next stage of his "grand strategy" wtih some guideposts which will assist him in deciding with what certainty the last point has been established.

In the above outline we may insert a step where statistical models can be of assistance.



This new portion may be carried out by a mathematical model which may approximate reality sufficiently closely to be very useful, particularly if the mathematical model allows for individual variation, i.e., a statistical model. Differential equations have often been used if individuals show little variation so that a representation by mean values is sufficient.

When the statistical model is used to assist the research there are additional feedbacks to the future working hypotheses both from estimates obtained incidental to the main study and from side investigations designed to aid in constructing the statistical model itself.

For a statistical model to be of real assistance it should fit as closely as possibly to the known characteristics of the experimental situation. The research worker will gain confidence in the statistician if the statistician is interested in knowing what type of observations he has because the type of observations available will affect the kinds of hypotheses and certainly the kinds of analyses (or models) which are appropriate. Classifications often used are:

Name	Type	Example
Nominal scale	Identification	White or black
Ordinal scale	Order	Test tubes ordered on color intensity
Interval scale	Intervals are equated	(Most common measurements)
Ratio scale	Zero defined	Distance

How Is a Statistical Model (Including Design) Specified?

Type of population (whether individual or measurements)

Sampling method

Definition of measurement

Hypothesis

What is "different"

What is known from previous experiments (may specify some parameters for this experiment)

Form of distribution of observations (or at least kind)

What next step is anticipated in line of research (estimate additional parameters for future design)

Statement of risks

α risk of rejecting hypothesis when true

β risk of accepting hypothesis when difference = d

Estimate of sample size (or sequential plan)

Expected troubles and controls or safeguards

It is well for the research worker to know what kind of problems a statistician considers simple and what kind are more difficult.

A. Examples of simple problems

<u>One variable</u>

Does a change occur in one measurement when a treatment is given?

Do two treatments differ in their effect on one characteristic of a patient?

Do several treatments differ in their effect on one characteristic of a patient?

Two variables

Are two measurements associated in the same way for two types of material?

B. Examples of problems less simple

Which one of 7 treatments is really different from the rest?

Does a change occur when a treatment is given when change may be reflected in any of several variables or combinations of them?

Which group of 4 or 5 measurements out of 28 measurements is best for assigning an individual to Group A and Group B?

Example: from amino acid measurements assign to group with or without hepatitis

Example: from physical and social variables assign mentally defective children to groups according to ability to profit from training

If a variable y is affected by changes in any of four x variables what is a good strategy for selecting the particular x values to optimize y?

The man developing the statistical theory will seek answers for the above questions, answers which will specify the appropriate model.

What Sort of Models Will the Statisitcian Give the Research Worker?

Most problems may be attacked by certain general statistical models. Or, a new model may be developed which follows more precisely the assumptions of the particular experiment under investigation.

The situation is somewhat like the tailor who has a shop with a number of ready-made suits. They can be used with little delay and sometimes slight alteration may be made. A tailor-made suit may require a great deal of time to construct, perhaps by many tailors and will certainly cost more. The ready-made suit will not fit everyone, but may fit a great many fairly well-some, very well.

What are some of the ready-made suits?

Scale	Statistics
Nominal	Binomial, multinomial, X^2
Ordinal	Order statistics, non-parametric statistics, median percentiles
Interval	Mean, standard deviation, standard test, t-tests, correlation, regression, analysis of variance
Ratio	Little different from interval scale except some parameters may be specified and not estimated

For analysis of observations on the interval scale most ready-made suits assume normality. How does one think about satisfying the assumption of normality?

a) Is it known?

b) Is it to be verified?

c) Will it be produced? (transformation--including averaging) (here we change shape of the man to fit the suit; he may have to wear a transformation before the suit will fit)

d) Will we show special caution in conclusions?

e) Will the assumption be avoided, e.g., by use of non-parametric methods? (this may be like using a suit that is too big--it may not pinch, but it may not fit closely anywhere)

f) Will a theory be developed for the appropriate distribution?

Another assumption which is made by many tailor made suits for analyzing observations on the interval scale is homoscedasticity (the assumption of equal variances). The same considerations can be listed for satisfying this assumption that were mentioned in the discussion of normality.

The research worker may be interested to know the criteria statisticians use in constructing models.

There is the goal of efficiency. Minimization of the number of observations for fixed risks α and β or minimization of one risk when the other risk and the sample size is fixed.

There is the goal of unbiasedness or accuracy. A statistical measure should be correct on the average.

There is the goal of minimum variance or precision. A statisitcal measure should have minimum variability as an estimate.

There is the goal of maximum power in the test of a statistical hypothesis. If the hypothesis is not true we should have a good chance of discovering this fact.

We must also tell the research worker that the particular statistic which will satisfy these goals depends on the type of observations which he makes. There is no universal answer.

It may be news to the person contacting the statistician for purposes of sanctification that the field of statistics also concerns itself with the construction of designs for the experimental attack itself. There are some answers to the questions:

Which cases should I select?

Which dose should be given next?

Which variables should I measure?

What combinations of treatments should be investigated together?

Do we make more measurements with no change in conditions or observe under more different conditions?

Some of these questions may come under the heading of the strategy of replication.

A chemist is quoted as follows, "I don't believe in replication. If you measure it once you know what you've got. If you measure it again and don't get the same answer, you don't know where you are."

Replication is often considered to occur only when several observations are made under identical conditions. The use of appropriately balanced designs can yield the advantages of replication at the same time one does experimentation over a wider range of other variables.

If we consider the example of investigating the effect of three variables at each of three levels, we make 27 observations all under different conditions if an observation is made for each combination of levels. But, since the experiment is performed in a balanced way we investigate the effects of changes in several variables simultaneously and can investigate the individual effects of each variable alone, the interaction of one variable with another and estimate the replication or measurement error. The randomized blocks and factorial designs can be used to advantage in medical research. However, we can also ask:

Is it essential to measure all possible combinations?

If not all are required we can perhaps suggest a carefully balanced subset which will still provide answers to the important questions. Such designs are latin squares or other fractional factorials.

My comments this evening are not intended to be comprehensive but only to indicate by examples the importance of both the statistician and the research worker continually educating each other and to list a few of the important points about which education should take place.

I wish to close with two comments which may be classified as philosophical.

First, there is a great concern in medicine for arriving at conclusions which will state a cause and effect relationship rather than an association. I have found it necessary to offer the information that statistical analyses in general only demonstrate association.

Second, the comparatively recent attention to Type II error in statistics may be at fault for its neglect in many scientific reports at the present time. Its continued neglect, however, may be tied to the notion of conservatism since the Type I or -error is controlled. In diagnostic situations this is often the risk of challenge to authority. There seems to be less interest in the risk of continued acceptance of authority which is not correct.

SAMPLING IN BIOLOGICAL POPULATIONS

D. B. DeLury
University of Toronto

It is not my wish to burden you with the practical details of methods that are used to estimate the vital statistics of biological populations. I propose only to skim lightly over some of the more interesting methods and to take advantage of the occasion to preach a sermon on the sin of non-randomness.

We have today an elaborate, well-developed Theory of Sampling, aimed chiefly at human populations and the things they do. The place where we get a grip on these problems is the fact that these populations are fixed geographically or in some other way that provides a basis for effective stratification. Furthermore, these populations have the property that we can, in principle at least, get at all the individuals in them and therefore we have a basis for a positive randomization procedure to select samples within strata.

We have also today a well filled-out discipline called the Design of Experiments, and this too has meaning only in circumstances in which effective randomization is possible.

Not everybody today is as convinced of the importance of randomization as he should be and consequently some of our investigations fall short on the score of randomness. Perhaps a look at the antics that people engage in when randomness is not possible will point up the essential role that randomness plays.

Some biological problems, of course, fall nicely within the scope of standard sampling procedures, e.g., a study of a population of nesting birds, or a beaver population, even though it might be difficult enough to carry them out. These things have, in fact, been done a few times, but the job of carrying out the dictates of randomness is formidable indeed.

It is not questions of this kind that I want to talk about, but those in which random selection of samples is truly impossible. I shall speak particularly of sampling populations of fish as, perhaps, the most striking instance of this. The impossibility of doing anything positive to ensure randomness in samples of fish is obvious enough, and those negative steps we might take to avoid the most unpleasant consequences of non-randomness are unknown or are known only in a qualitive way. Furthermore, those features of the populations which can upset our procedures are known only qualitatively.

Fish are generally, though not always, highly mobile. They tend to stratify according to age or size and in other ways as well. Every method of capture we have devised is biased with respect to size and doubtless other features too. Indeed, the conviction is growing that the probability that a fish will be caught by any particular device varies widely from one individual fish to another.

In any event, let us look at some of the things people do, to try to estimate the size of a population of fishes, keeping in mind the difficulties I have been talking about. Nothing here is at all new; this is simply an account of things that have been done.

TAGGING. One of the older tricks is to catch a sample of fish, put tags on them or otherwise mutilate them so they can be recognized, release them and watch for their reappearance in subsequent catches. Then, if we can assume that the proportion of marked fish in the catch is equal to the proportion in the population, apart from sampling error, we get at once an estimate of the size of the population. This device goes back to Laplace, I believe, but he did not use it on fish.

To put this into symbols, using t to denote the time since the tagged specimens were released:

	Tagged	Total
Population	X_t	N_t
Sample	x_t	n_t

$$\frac{x_t}{n_t} = \frac{X_t}{N_t} = \frac{X_o}{N_o} \quad , \qquad N_o = \frac{n_t X_o}{x_t} \quad ,$$

valid as long as any depletion affects tagged and untagged equally.

While there are many things one can think of which would upset this estimate, the weakest spot is surely the supposition that the proportions in the sample and in the population are equal -- i.e.,

$$\mathcal{E} \quad \frac{x_t}{n_t} = \frac{X_t}{N_t} \quad ,$$

or to put it otherwise, tagged and untagged are equally catchable. This is the kind of thing we look to randomness to ensure, the kind of thing we expect to be met a priori through the way we select our samples. If we were dealing with beads in a box, instead of fish, we would mix them thoroughly after introducing the marked beads and then draw out our samples. With fishes, what can we do? Perhaps not very much, but we are coming to know some of the pitfalls. To illustrate one of the common ones, I can tell you about a tagging experiment carried out some years ago to ascertain the number of black bass in a rather large lake. The bass were captured in traps, put here and there around the shore and left there the whole summer. Many bass were tagged and recaptures were numerous but the estimate of the total number of bass, based on these recaptures, was something like 500, a completely fatuous figure. Now it happened that good records had been obtained from the anglers in the lake, both numbers caught and numbers tagged. An estimate based on these records was around 30,000. What was happening here? A run through the records disclosed that, with

practically no exceptions, every bass recaptured was taken in the same trap in which he was first caught. We can see, then, what happened. As far as the traps were concerned certain individuals were highly exposed to recapture, over and over, while others had little chance of capture at all. The same effect has been seen in other populations, for perhaps different reasons. One thing to be feared, then, is that fish captured in any particular manner are more likely than the others to be captured again in the same manner -- i.e., tagged and untagged are not equally catchable. This is particularly dangerous when repeated recaptures are used. It seems prudent, then, to recapture by a method that is different from the one used in placing the tags or, better still, use several methods for both and keep such records as are needed to keep track of the methods by which each fish is captured.

Whatever we may do, however, the tagging procedure provides no real check on the crucial assumption that tagged and untagged are equally likely to be caught. Information of another kind is needed if we are to do so.

Tagging procedures have been extended far beyond the simple one described here. We can, for example, tag and sample simultaneously in various ways. Such procedures, in which we can no longer treat as constant the proportion tagged in the population, will be distorted by the operation of an appreciable mortality during the sampling period, which adds another source of uncertainty. The fact that they are so distorted means, of course, that they contain information about mortality and a few schemes for extracting estimates of the mortality rate from them have been proposed. For most part, they depend on repeated recaptures and, for this reason, our inability to sample randomly strikes them particularly hard. We can side-step this dependence on repeated recaptures, but the methods we must then use are very weak.

Perhaps you will permit me to take off here on another tack to speak of a question that has vexed me for some time, because I see no good way of getting a grip on it. It is a point, though, that may have some importance outside the immediate context. One of these tagging plans furnishes a good vehicle for the discussion.

Let us suppose that we tag and sample simultaneously, perhaps by tagging and releasing all untagged members of each sample and releasing also those already bearing tags. To keep the discussion free form difficulties that are here irrelevent, let us say that the size of the population is constant throughout the sampling period, i.e., no mortality, immigration and so on. Let N stand for the size. Then, using the notation written down earlier, and making the reasonable assumption that each sample is small compared to the size of the population; i.e., the sampling is effectively binomial, we can write the probability of getting x_t tags in a sample of n_t:

$$\binom{n_t}{x_t}\left(\frac{X_t}{N}\right)^{x_t}\left(1-\frac{X_t}{N}\right)^{n_t-x_t} = f(t), \text{ say, and the likelihood}$$

is

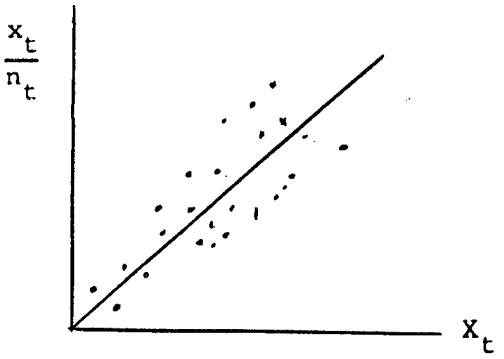
$$L = \prod f(t).$$

A direct maximum-likelihood calculation yields, after some algebraic rearrangement, the estimating equation

$$\sum \frac{x_t - n_t X_t/N}{1 - X_t/N} = 0 .$$

This equation can be solved for N by numerical methods, but this does not concern us here.

We can take another view of this question. Since $\mathcal{E}\frac{x_t}{n_t} = \frac{X_t}{N}$, we might think of plotting $\frac{x_t}{n_t}$ against X_t and fitting by least squares a line passing through the origin. The slope of this line estimates $\frac{1}{N}$.



Now, granting randomness in the sampling, the only error is the binomial error and the proper weights are given by the standard binomial formula

$$\frac{n_t}{\frac{X_t}{N}\left(1 - \frac{X_t}{N}\right)} .$$

Using these weights, the least squares equation can be rearranged into precisely the M.L. Equation, as might, perhaps, be expected. The point here is, then, that the M.L. estimate uses as weights functions of the population proportions $\frac{X_t}{N}$. Now if the sampling is, in fact, random, this is presumably the most effective weighting. However, in these circumstances, it seems reasonable to question this weighting, because, when any sample is taken, the proportion available to it may be grossly different from $\frac{X_t}{N}$. It seems to me, therefore, safer to weight by sample size, n_t, which incidentally leads to a simple estimating equation. The general question, then, is: when our sampling may fall short with respect to randomness, are our so-called efficient statistics even as good as other, less efficient ones? I do not think any general answer is possible here. The question itself if not precise and cannot be accurately formulated. I have, however, compared the two approaches outlined above on artificial data in which randomness is assured and can perceive no persistent difference between them.

One could spend several hours on ramifications of the tagging method, but I must pass on to a procedure of a different kind.

CATCH-EFFORT. This method rests on a common observation that, as a population becomes depleted, the return from a given amount of sampling effort declines. In order to make quantitative use of this fact, some assumptions are necessary and data of a specific kind must be collected. Let us say, then, that for each of a number of consecutive intervals, we have a record of the catch - $c(t)$ and the effort expended - $e(t)$. $e(t)$ might be measured in net-nights, boat-hours, etc. Then, we can calculate for each interval the catch per unit of effort, $C(t) = c(t)/e(t)$ (t refers to the t^{th} interval). We can calculate also the accumulated catch and the total effort expended up to the t^{th} interval.

$$K(t) = c(1) + c(2) + \ldots + c(t-1),$$
$$E(t) = e(1) + e(2) + \ldots + e(t-1).$$

Now, let us make the simplest assumptions we can think of, say that the population is closed and that a unit of effort takes always the same fraction k of the population. k has been called by some the "rate of fishing" and by others the "catchability" -- in any event, k is measured in $(\text{units of effort})^{-1}$.

We have, now, from the definitions and assumptions: $c(t) = k\, e(t) N(t)$, $N(t)$ being the size of the population at time t, or, dividing by $e(t)$:

$$C(t) = kN(t) \quad .$$

Also, because fish can leave the population only through being caught,

$$N(t) = N(1) - K(t) \quad .$$

Hence
$$C(t) = kN(1) - kK(t) \quad .$$

Now C and K are observable quantities. We can plot them, and if reasonable straightness results we can fit the line and so estimate $kN(1)$ and k, hence $N(1)$.

From the same assumptions, we can derive another relation,

$$\log C(t) = \log \left[kN(1)\right] - kE(t).$$

While these two relations are equivalent in a mathematical way, as soon as they are embedded in a statistical setting, they show up quite different. The first has only to do with the conditional behavior of $C(t)$ given $K(t)$, which depends only on $K(t)$, whereas the second depends on the whole stochastic process up to the t^{th} interval. As far as I know, this has never been elucidated and we know less about how to fit it than we do for the first.

We may properly be concerned about the rather restricted assumptions under which these relations have been derived and indeed we may be sure that frequently they will be violated. On the other hand, there is a certain amount that can be done to ensure that they will hold reasonably

well. The one assumption that is largely outside our control is the one which asserts that a unit of effort always takes the same fraction of the population -- i.e., that the catchability is constant throughout the sampling period. Here again is an assumption that we would look to randomness to take care of. Everyone knows that this parameter is bound to fluctuate, perhaps widely, from day to day, but this is not as serious as the possibility of persistent trends. We may expect trends in the catchability to distort our graphs out of straightness, but it must be granted that day-to-day fluctuations are usually large enough to make decisions on this question rather difficult. Furthermore, other failures in our assumptions can produce curvature.

On the whole, then, we find ourselves with two methods, each dependent on a crucial assumption which cannot be tested with the kind of data usually collected in applying this method. Each of them is the kind of thing that we usually look to random sampling to take care of. In this case, the best we can hope for is to try to find independent checks on the critical assumptions.

A little reflection will show that each of these methods contains information that can be used to test the assumption on which the other depends, provided that the two methods, tagging and catch-effort, are applied to the population simultaneously.

The catch-effort method requires that catchability remain constant and the tagged portion of the population provides a population of known size, for which direct estimates of the catchability can be made from the sequence of returns. These estimates can then be inspected for trends. The tagging method, on the other hand, requires that tagged and untagged be equally catchable. A catch-effort analysis, conducted separately on the tagged and untagged parts of the population, puts these two catchabilities directly in evidence, so that a comparison can be made.

Investigations of this kind are necessarily elaborate and expensive and, as far as I know, this combined tagging and catch-effort study has not often been attempted. Tagging alone has been used many times, but it is likely that most of the population estimates so reached are seriously wrong. The catch-effort method is newer and has been used much less. We need more experience with it, but we <u>can</u> say that, in some instances at least, it works pretty well. When the sampling is extensive enough and intense enough to bring about heavy depletion in a rather short time, we may expect the catch-effort method to function fairly well. This does happen not infrequently in commercial fisheries and some sports fisheries.

I have not spoken of the elegant and theoretically powerful methods developed during the past few years by a number of people, P. H. Leslie and D. G. Chapman, to mention two. These methods depend on tagging and in some instances, on catch-effort as well, and they purport to yield estimates of all the vital statistics of a population, birth and death rates and so on, as well as population size. Not only do these methods rest heavily on random selection of samples, depending as they do on <u>repeated</u> recaptures, but they also involve actuarial notions such as <u>mortality rate</u>, presumably

constant and stable. Now these notions work well enough in human populations, because we have pretty well eliminated the catastrophe from our system of causes. Likewise, actuarial methods, i.e., methods based on mortality tables, may well be applicable in some controlled populations, e.g., laboratory populations. In natural populations, however, the most prominent effects are environmental, fluctuations are wide and capricious, masking entirely the built-in cause-and-effect mechanisms without which no population can survive. Only when the populations are extraordinarily dense or sparse do these mechanisms play a predominant role. For these reasons, I think we have to keep our sights fairly low in dealing with natural populations; in particular, we should not use methods which depend on the kind of stability we have become used to in human populations.

THE APPLICATION OF FRACTIONAL FACTORIALS IN MISSILE TEST PROGRAMS

Paul C. Cox
Ordnance Mission, White Sands Missile Range

I wish to begin by acknowledging the assistance I have received from Dr. Boyd Harshbarger in developing and applying the specific fractional factorial which I plan to discuss. I understand he had an entire class at VPI work with this design, and the fact that it is a design which is presently being used for a specific missile test program proved to be both stimulating and motivating to the class.

The statistical designing of a missile test plan is usually plagued with numerous serious problems. These include: (1) There are usually a large number of different types of treatments and at several levels which must be evaluated; (2) the sample size is almost always quite small; (3) there are quite often, physical limitations which will place severe restrictions on the design; (4) the test must be designed with the realization that the statistical evaluation is but one of many requirements to be satistied by the test; (5) most missile test plans require several changes after testing has started; and (6) there will invariably be a few test failures which will either require substitute firings or will result in a loss of data.

I will restrict my discussion to the first two of these problems the large number of treatments required and the restricted sample size. Because of these characteristics, fractional factorials can sometimes be a useful tool in designing missile test plans. Fractional factorials are valuable because they provide for a reduction in sample size. This reduction is bought at a price and the price should be carefully studied before using a particular fractional factorial for a specific missile test plan. Actually, we at White Sands Missile Range consider fractional factorials frequently, but because of the limitations, use them rarely.

I will illustrate these points by discussing a problem we were recently faced with and how a fractional factorial appeared to be the best solution of the design problem.

It was required that a missile system be flight tested to determine its effectiveness under the following combinations of test conditions:

Type of Warhead	W_1	W_2	W_3
Propellent Temp.	T_1	T_2	T_3
Launcher-Target Range	R_1	R_2	R_3
Launcher Emplacement	E_1	E_2	E_3
Launcher	L_1	L_2	

(Two identical launchers).

It was decided that the missiles should be fired from the two launchers in pairs. In this way, the error mean square will contain only the

variability due to the missile system and no variability due to changes in weather, firing personnel, etc.

Ideally one should use a factorial design with a minimum of four rounds tested under every set of conditions. (One pair at one time and the second pair at some other time). This would require a total of 3x3x3x3x4 = 324 rounds; a figure which is entirely unrealistic.

Since a factorial design could not be used, the obvious procedure was to study the physical characteristics of the test and try to develop a fractional factorial which would be suitable. It was determined from the study of the physical characteristics of both the system and the test that emplacements would probably not interact with any of the other test conditions. Consequently a fractional factorial was designed by setting up R, W and T as a factorial design; then for R_1, E was introduced in the form of a Latin Square and similarly for R_2 and R_3 except the rows were permuted. (1) The same design is used for both L_1 and L_2 since the rounds are fired in pairs from the two launchers. The design is given in table 1. This is referred to as a 3^4 confounded fractional factorial in blocks of 27 units for which 1/3 of a replicate is given, and it requires a total of 54 rounds plus spares, a figure which is realistic in this particular case. A design similar to this one is plan No. 6A.18, P 290, Cochran and Cox. (2)

		L_1 and L_2		
		W_1	W_2	W_3
R_1	T_1	E_1	E_3	E_2
	T_2	E_3	E_2	E_1
	T_3	E_2	E_1	E_3
R_2	T_1	E_2	E_1	E_3
	T_2	E_1	E_3	E_2
	T_3	E_3	E_2	E_1
R_3	T_1	E_3	E_2	E_1
	T_2	E_2	E_1	E_3
	T_3	E_1	E_3	E_2

Table 1. A 3^4, 1/3 replicate, confounded Fractional Factorial in Blocks of 27 units.

(1) It would have been possible to use Graeco-Latin Squares, and thus test some other treatment at 3 levels, providing certain assumptions could be made about this new treatment.

(2) "Experimental Designs," Second Edition, W. G. Cochran and G. M. Cox, 1957, John Wiley and Sons, Inc., New York

The precautions to be observed in this design are: (1) The RT, TW, and RW interactions are aliased with the E interactions; (2) The RWT interaction is aliased with the E Main Effect; and (3) It is doubtful if a suitable comparison can be made between the variability in performance due to the missile and the variability due to the missile plus metric conditions.

These difficulties were all studied carefully and it was agreed that the design in figure one was appropriate as far as our problem is concerned. It was felt that the likelihood of E interacting with R, W or T is negligible, and that the RWT interaction is probably negligible, also. The remainder of the discussion is based upon the hypothesis that the above assumptions are all correct.

Table 2 shows two analysis of variance tables. One is for the fractional factorial design described in table one and the other is for a 3^4 factorial with 2 replications.

Sources of Variance	Fractional Factorial	Factorial degrees of Freedom
W	2	2
T	2	2
R	2	2
E	(WRT) 2	2
WR	4	4
WT	4	4
RT	4	4
WE		4
RE		4
TE		4
WRT		8
WRE		8
TRE		8
WTE		8
WRTE		16
Interactions associated with E	6	
Error	27	81
Total	53	161
Sample Size	54	162

Table 2. Degrees of freedom associated with a 3^4 factorial with 2 replications and a 3^4 fractional factorial showing 1/3 of a replicate (repeated).

From table 2 it is clear that, if the assumption that the E interactions and the WRT interactions are negligible, we are buying almost as much from the fractional factorial with 54 rounds as from the complete factorial with 162 rounds. It is true the degrees of freedom for error are 27 as compared to 81 for the complete factorial, but the increase in power which results from using 81 rather than 27 degrees of freedom is usually unimportant.

The computation of sums of squares is extremely simple for this design. The procedure is the same as for a complete factorial design except the error term must be computed by taking one half the sum of squares of the difference between pairs. Then the term which we describe as interactions associated with E is obtained by subtraction.

This design will be illustrated by the data given in table 3 and the analysis of variance is given in table 4. The data of table 3 is fictitious, such data was chosen for two reasons: (1) to keep this presentation unclassified; and (2) the actual study is not far enough along to provide actual data. The data listed in table 3 is radial miss distance which has been transformed in order that an analysis of variance would be appropriate. The important thing to note is that the smaller the value the better the weapon has performed.

			L_1			L_2	
		W_1	W_2	W_3	W_1	W_2	W_3
R_1	T_1	36 E_1	38 E_3	31 E_2	36 E_1	46 E_3	15 E_2
	T_2	14 E_3	15 E_2	20 E_1	20 E_3	24 E_2	35 E_1
	T_3	32 E_2	33 E_1	25 E_3	29 E_2	26 E_1	24 E_3
R_2	T_1	21 E_2	29 E_1	32 E_3	20 E_2	26 E_1	22 E_3
	T_2	8 E_1	6 E_3	16 E_2	4 E_1	13 E_3	11 E_2
	T_3	28 E_3	14 E_2	13 E_1	22 E_3	12 E_2	22 E_1
R_3	T_1	45 E_3	33 E_2	32 E_1	42 E_3	37 E_2	27 E_1
	T_2	16 E_2	33 E_1	32 E_3	15 E_2	34 E_1	32 E_3
	T_3	33 E_1	45 E_3	25 E_2	40 E_1	42 E_3	35 E_2

Table 3. Results of a firing test using a 3^4 fractional factorial and two launchers.

Sources of Variation	S.S.	D of F	M.S.	F Ratio
W	100	2	50	2.17
T	1409	2	704	30.61 **
R	2223	2	1111	48.30 **
E (WRT)	466	2	233	10.13 **
WT	695	4	174	7.56 **
WR	167	4	42	1.83
TR	109	4	27	1.17
Due to E Interaction	156	6	26	1.13
Error	626	27	23	
Total	5951	53		

Table 4. Analysis of Variance for data in Table 3. (** Indicates significance at the 1 percent level).

Table 4 indicates that all main effects except for warhead type, have a significant influence upon missile accuracy, with range having the greatest influence of any. WT is the only significant second order interaction. At this time, it is desirable to investigate the mean square which is attributed to the various E interactions. If this were not small, one might have some doubt about whether the assumption concerning the E interactions being negligible was really sound. If we were absolutely certain that E interactions are not possible, this mean square term could be used for another purpose. If there had only been one replication (27 rounds) it would have been necessary to use this term as the error term. But since the rounds were fired in pairs we may now consider that the mean square attributed to E interactions is an estimate of the variability which is due both to the missile and to the day to day variation, while the mean square for error term is an estimate of the missile variability after stripping out the day by day variation. By comparison it is seen that the difference between the two estimates is negligible from whence it might be implied that day by day variation appears to be well under control.

Returning to the results of table 4, the overall mean is found to be 26.22 and the mean values for the various levels for the main effects are given by table 5 below.

Treatments \ Levels	1	2	3
W	25.61	28.11	24.94
T	31.55	19.33	27.78
R	27.72	17.72	33.22
E	27.05	22.28	29.33

Table 5. Mean Values for Each Level of W, T, R and E.

From table 5, it appears that warhead No. 3 results in the smallest while warhead No. 2 causes the greatest miss distance. However, the effects of warheads are not significant and there is no reason to believe that one warhead will cause a larger miss distance than another. Ambient temperature results in the smallest while low temperatures cause the greatest mean miss distance. Medium ranges have the smallest, while long ranges have the greatest mean miss distance. Finally it may be seen that launchers emplaced on level ground will result in the smallest mean miss distance while an emplacement on the fore side of a hill will result in the greatest mean miss distance. It would be very much in order to study the WT interactions, but this will be omitted, largely because the main effect W is not significant.

Conclusions: When testing missile systems it is usually the case that many levels of treatments must be studied from the data obtained from a limited sample size. Fractional factorials are frequently a very useful tool for designing such tests. There are many pitfalls to watch for when using fractional factorials, but many times this technique appears to give results nearly as good as those obtained from a complete factorial and by using a much smaller sample size.

THE DESIGN & REDESIGN OF AN EXPERIMENT

C. W. Mullis
Integrated Range Mission, White Sands Missile Range

INTRODUCTION. In May 1958, White Sands Missile Range undertook an evaluation which was particularly amenable to optimization of the experiment. An experiment was designed and data collection began. Unanticipated field problems required approximately four times the effort predicted in order to fill enough points in the design matrix to permit reasonable analyses. The end result was that an experiment expected to take six months, extended over a period of sixteen months and is just now nearing completion. In view of these circumstances, this paper might better be titled by using the often quoted expression "The best laid plans of mice and men often go astray."

THE PROBLEM. White Sands has been employing cinetheodolites manufactured by Askania-Werke A. G. since the range was established in 1945. Today there are approximately sixty instruments in regular use. In recent years a new instrument has appeared on the market which is purported to represent the state-of-the-art in cinetheodolite type instruments. This instrument, manufactured by Contraves A. G., Zurich, Switzerland is known as the Contraves EOTS. During 1958, the J. W. Fecker Division of the American Optical Company became the United States distributor and was desirous of obtaining information on the comparable performance of their new product and existing equipment in the field. They also wished to demonstrate that the Contraves was dynamically accurate to better than 5 sec of arc (one part in 250,000). Due to the potential market, the varied nature of the missile firing workload (approximately ten missiles of varying types fired each work day) and a unique capability to install instruments in dual installations side by side, Fecker proposed to White Sands that evaluation of comparison of the Contraves instrument to the existing Askania instruments be performed. Previously the only available comparison was Contraves at Eglin to Askania at White Sands or similar cases which left much room to challenge validity. White Sands welcomed the opportunity to settle the argument, and obtain first hand information on the new instrument.

CONSIDERATIONS. By utilizing the dual installations, operating on the same missions from the same timing distribution and control network, reading and reducing the data in the same plant, and comparing data taken at the same instant in time, it was felt that a valid comparison of the instruments in question would be forth-coming. The design of the experiment then centered around the features of the instruments and their deployment.

The first question was "How many to use." Fecker proposed to furnish two Contraves. It is a well known fact that although two station triangulation meets the mathematical requirement for a solution, the accuracy of the final computed data increases as the number of stations is increased. Consequently use of three instruments for each system was agreed upon.

The next question was "where to put them." Figure 1 shows the deployment of the existing Askanias. The nature of the missions can be summarized as follows: Ballistic type missiles are fired from the Small Missile Range near "N" Station, and from the launching strip extending

eastward from the Army Missile Test Center. Those missiles may impact in the 30, 50, 70 or 90 mile impact areas. Air to air missions are conducted above the four major impact areas. Surface to air missiles are launched from the launching strip and may intercept anywhere from a few miles north of the launch point to the northern boundary of the range. It was desired to locate the instruments such that as many of the various types of missions as possible could be included in the experiment.

Another factor influencing the placement of the instruments was intersection geometry. Needless to say, the accuracy of final data is quite sensitive to the angle of intersection of the lines of sight from the instruments to the target. Tracking capabilities should be taxed to the maximum but the capability of the instruments was not to be exceeded. An arbitration of these factors resulted in the deployment shown in Figure 2. This deployment had good geometry for at least two of the instruments on missiles launched from any launch area. Data could be obtained on all but air to air missions. The instruments experience varying modes of tracking severity depending on the launch point. All have wire lines for communication and timing distribution and they are all "close in."

The next consideration was the physical installation of the instruments. Figure 3 shows an actual installation with the Askania on the left and the Contraves on the right. The Askania is mounted on a hydraulic hoist which elevates it through a hole in the roof of the monolithic concrete building. The kinematic platform on top of the hoist is then rotated, locked into the building roof, and the hoist lowered leaving the instrument "sitting" on the building. Obviously, the Contraves would not pass through the hole. Therefore, a one inch thick steel plate was fabricated to cover the hole and the Contraves mounted to the plate. In essence then both instruments were using the same pedestal, namely the building.

Since all film was to be automatically processed by the high speed continuous processing machines, it was felt that no special consideration should be given this phase of the test.

When the film reached the data reduction portion of the system, other questions arose. For instance, what about human error in film measurement? In order to minimize this it was decided to read each target board frame five times and each data from three times and use the averages thus obtained as the reading.

Registration in the reader was another questionable area. Since the Contraves mechanism is pin registered and the Askania mechanism is not, it was decided to check the registration for both by resetting on each frame. In the Contraves, the frame is rotated through the elevation angle. The center of the recorded frame then moves according to the eccentricity of the elevation axis. Therefore the center of the frame was rechecked for each frame. For the Askania, the fiducials were checked as usual, each frame.

CALIBRATIONS. Calibrations are the measurements made on the instrument to determine the bias errors which exist in its various parts. A correction thus obtained is applied in the data reduction process so that biases do not

appear as errors in the end result data. These measurements may occur at very infrequent intervals or may be made each time the instrument is operated. Figure 4 shows the calibrations performed on the two types of instruments for this test. Lens sag does not exist for the Contraves and circle eccentricities are eliminated by diametrically opposed scale readings. However, these were measured and accounted for in the Askania. Although lens distortion was measured for both instruments and for the projection lenses of the film readers, it was not necessary to use it since the tracking was good. That is to say the target was always near enough the center of the frame that distortion could be neglected. To insure that no appreciable error was contributed by the film reading machines, the measuring cross hair digitizers were calibrated. The measurements thus obtained were used to correct the final data which were used for computation.

HYPOTHESES. As previously stated, Fecker desired to prove that the Contraves was accurate to 5 sec of arc or better. The Government wished to compare the performance of the two systems. An attempt was therefore made to design an experiment which would achieve both goals. The null hypotheses were stated as follows:

TEST 1

Null Hypothesis

The Contraves EOTS cinetheodolite does not exhibit smaller random errors than the Askania Kth 53 cinetheodolite.

Alternate Hypothesis 1

The Contraves EOTS cinetheodolite exhibits random errors less than 1/2 the value of the random errors of the Askania Kth 53 cinetheodolite.

Alternate Hypothesis 2

The Contraves EOTS cinetheodolite exhibits random errors less than 1/4 the value of the random errors of the Askania Kth 53 cinetheodolite.

TEST 2

Null Hypothesis

The Contraves EOTS cinetheodolite system (three stations) does not exhibit smaller random errors than the Askania Kth 53 cinetheodolite system (three stations).

Alternate Hypothesis 1

The Contraves EOTS cinetheodolite system exhibits random errors less than 1/2 the value of the random errors of the Askania Kth 53 cinetheodolite system.

Alternate Hypothesis 2

The Contraves EOTS cinetheodolite system exhitits random errors less than 1/4 the value of the random errors of the Askania Kth 53 cinetheodolite system.

These hypotheses are based on the Askania as a standard. By testing the Askania random errors against, the 5 second figure, it was felt than an absolute value for the accuracy of the two systems would be obtained.

To test these hypotheses, two methods were to be employed. For Test 1 the Variate Differences 1 techniques was to be used. For Test 2 a statistical analysis of the residuals from a standard Davis solution 2, 3 was to be used. Test 1 would give a comparison of instrument precision, where Test 2 would give system accuracy including biases. Since, by the nature of the techniques Test 2 required considerably more data than Test 1, Test 2 governed the determination of the data to be collected. The experiment then resolved itself into a study of two variables, one containing six parameters, the other two parameters to be investigated on three levels. A further division occurs when it is considered that each system containing three instruments is treated as a single variable. The design matrix is shown in Figure 5.

DATA COLLECTION. From the design matrix, it was planned to collect data on eighteen missions. This was to be accomplished in one month to six weeks. In actuality data were collected over a period of three months on eighty seven missions. From this we were able to select ten missions on which all six instruments (three Askanias and three Contraves) functioned properly and collected sufficient data for analysis. This is not to say that the instruments are that unreliable. For proper application of the variate differences analysis it is necessary to have 100 consecutive points. Loss of a single point due to a condition of mistrack eliminated numerous records from consideration. Loss of points due to the missile passing behind clouds had the same effect. Another difficulty was the fact that although 100 consecutive points may have existed for the various instruments, they did not overlap enough in time to allow a valid comparison.

FIELD PROBLEMS. As data collection progressed operational observations lead to questioning of the assumptions. Primary among these was the rigidity of the building. A rather long term effect was observed in the tilting of the building due to differential heating. On one particular day this amounted to approximately 40 seconds of arc over a period of 6 hours. This was not considered to be significant in terms of the effect during a particular mission since the mislevel of the instrument is read before and after each mission. However, one observer noted that there was a sudden shift of approximately 10 seconds of arc which occurred in a period of a few minutes. The reason for this shift could not be explained.

As a result of the uncertainty in the stability of the support structure, a method was devised whereby the movement of the building could be measured during a mission. This consisted of checking the movement of the

instrument base against a pendulum mirror which was referenced to earth's gravity. [4] Only a few measurements were made but the results showed that no significant movement of the building occurred during the mission.

Another assumption which is made in the usual operation of the system is that the target boards which are used to determine the orientation and mislevel of the instrument are stable. The azimuth stability of the target boards was noted in several spot checks using the Contraves film information. In most cases, the correspondence was within 3.6 seconds of arc. The elevation position of the target boards was not checked as accurately as this since turbulence research at White Sands has indicated that there is much less vertical deterioration than horizontal deterioration due to turbulence. However, refraction in the vertical direction may be significant.

Of the three stations used in the field, one station, "N", did not "fit in" with the other two for either the Contraves or the Askania. This means that the line of sight from that station did not pass through the intersection of the lines of sight from the other two stations within the prescribed limits of the data reduction procedure. Thus far, no explanation has been found for the large discrepancy observed. Surveys have been rechecked. Timing distribution delays were measured to be less than 100 micro-seconds. All known possibilities have been exhausitvely checked with no answer being found.

ANALYSIS. Due to their simplicity, the variate differences calculations were started first. Although they have not yet been completed, a sample tabulation of some of the data from three missions is given in Figure 6. From this tabulation one can see the nature of the information which is being obtained. For instance one can make instrument to instrument comparisons in azimuth and/or elevation; compare the azimuth or elevation performance of like instruments; compare azimuth to elevation performance of a single instrument, etc. By performing a statistical analysis of the data for several missions, generalized statements relative to the characteristics of the instruments can be made. Once the data have been calculated for all the missions these analyses will be completed.

The analysis for Test 2 does not present so bright a picture. With the exception of one mission, it has been impossible to get the data from "N" station to work in a three station trajectory computation. Forced solutions have had residuals far beyond the rejection limits. On one mission the data would work in a two station solution with either of the other stations but three stations would not run. To date, this problem is still under consideration.

CONCLUSIONS. Time does not permit an excursion into the details of system improvements which have been effected as a result of the deficiencies discovered during this evaluation. One significant course of action as far as the experiment is concerned has been decided upon. Fecker is planning to take the raw data from some of the missions, read the data on their own equipment and perform an independent analysis. All the procedures will be similar except for the computer program, which will account for an assumption made in the Davis solution to linearize the equations.

White Sands Missile Range has proceeded with two station solutions using "C" and "T" stations and is preparing a statistical analysis of these data for Test II.

In summary, it may seem strange that this discussion has been presented to this group. The intent was to emphasize the physical problems encountered and the requirement for flexibility in the statistical design of experiments of this nature.

REFERENCES

1. M. G. Kendal, "The Advanced Theory of Statistics," Vol. II, P387-394, Hafner & Co., New York.

2. R. C. Davis, "Techniques for the Statistical Analysis of Cinetheodolite Data," NAVORD Rpt. 1299, NOTS 369, Naval Ordnance Test Station, China Lake, California, 22 March 1951.

3. F. P. Apostalas and J. B. Gose, "Askania Cinetheodolite Procedure Technical Memorandom 446, White Sands Missile Range, New Mexico, August 1957.

4. E. C. Schluter, "Cinetheodolite Dynamic Accuracy," Instrumentation Services Interim Report: April 1959, Measurements Division, Integrated Range Mission, White Sands Missile Range, New Mexico.

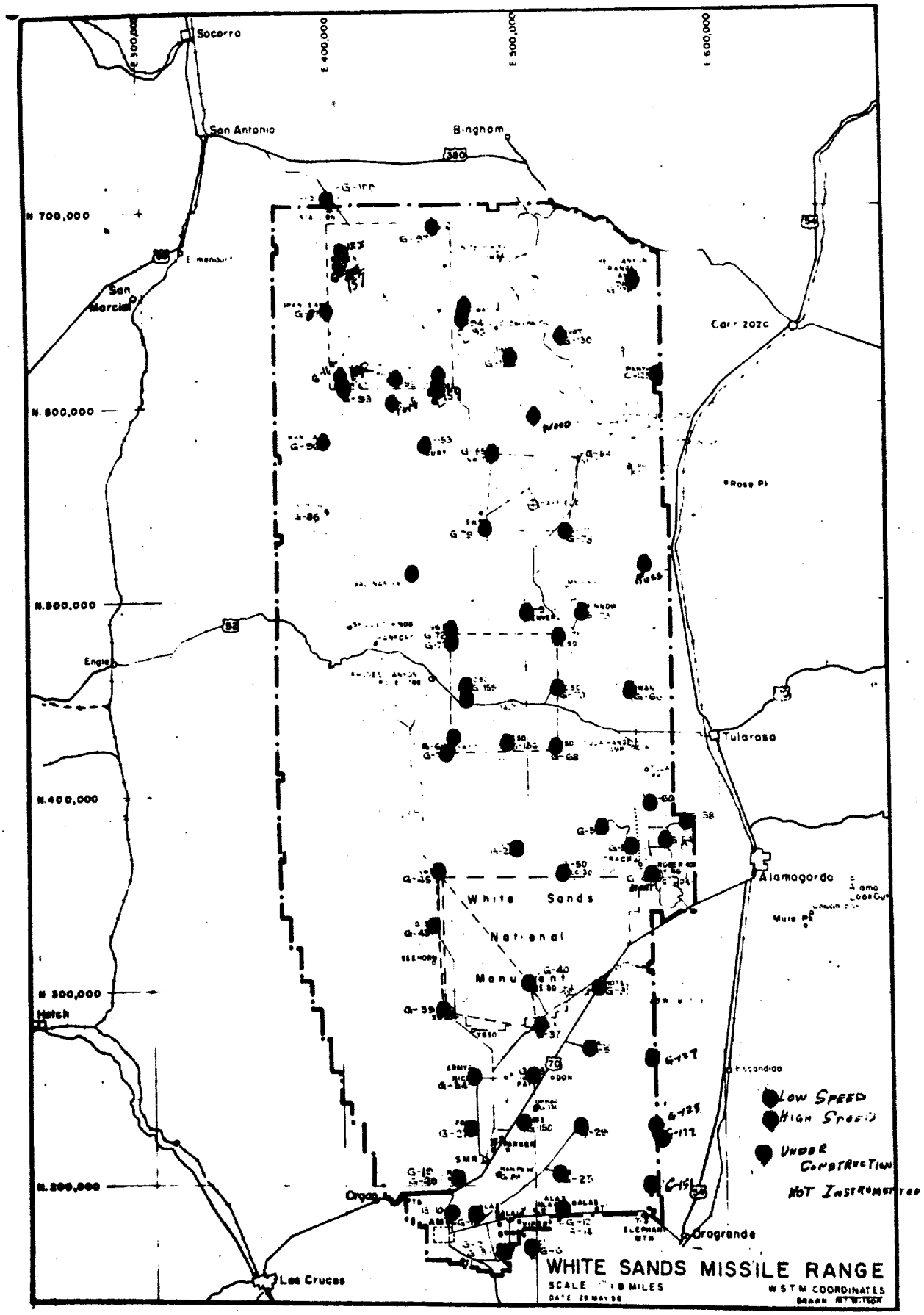


FIGURE 1. CINETHEODOLITE INSTRUMENTATION

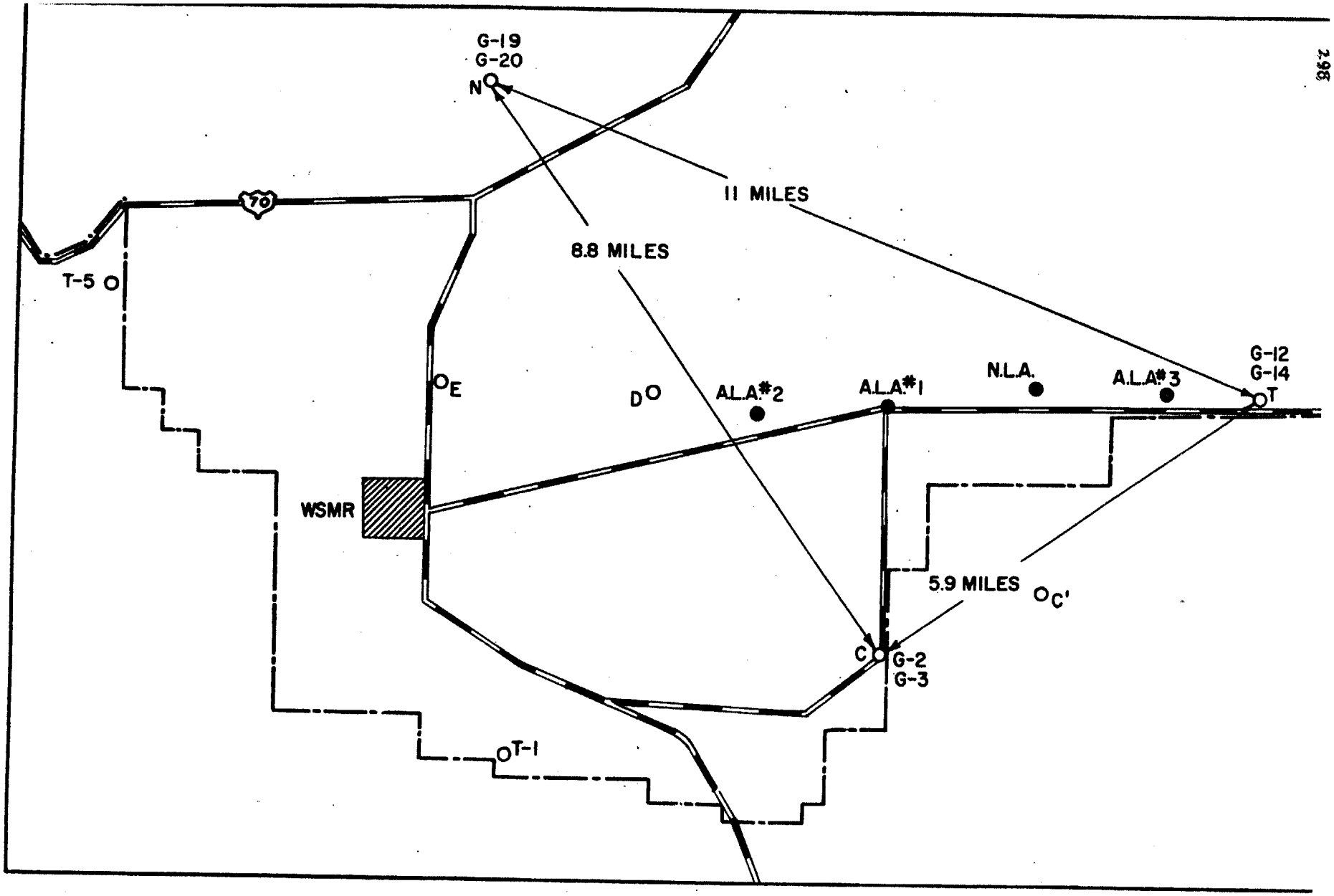


FIG. 2. INSTRUMENT DEPLOYMENT SHOWING THE RELATIONSHIP TO THE LAUNCHING AREAS. G-2, G-12, & G-19 ARE THE ASKANIAS. G-3, G-14, & G-20 ARE THE CONTRAVES.

	ASKANIA	CONTRAVES
TRUNION AXIS ERROR	X	X
COLLIMATION ERROR	X	X
MIS-LEVEL	X	X
LENS DISTORTION	X	X
LENS SAG	X	
CIRCLE ECCENTRICITY	X	

FIGURE 4 CALIBRATIONS

VARIABLE	PARAMETER	LEVEL		
		HIGH QE BALLISTIC	LOW QE BALLISTIC	MANOUVERED
ASKANIA	5/SEC B&W	X	X	X
	5/SEC COLOR	X	X	X
CONTRAVES	5/SEC B&W	X	X	X
	5/SEC COLOR	X	X	X
	10/SEC B&W	X	X	X
	10/SEC COLOR			
	20/SEC B&W	X	X	X
	20/SEC COLOR			

FIGURE 5 EXPERIMENT MATRIX

NOTE: EACH POINT IN THE MATRIX REPRESENTS FROM 1500 TO 6000 DATA POINTS.

COMPARED BIASIS (MINS)		MISSION NR.	VARIATE DIFFERENCES (DEGREES)											
STATION COMBINATIONS	INSTRUMENT BIAS σ_B^2		AZIMUTH		ELEVATION		AZIMUTH		ELEVATION		AZIMUTH		ELEVATION	
			G-2	G-3	G-2	G-3	G-12	G-14	G-12	G-14	G-19	G-20	G-19	G-20
		040001		.000039		.000743		.066414		.003246				
						22		9		4				
						19		5		3				
						18		5		3				
						17		5		3				
						15		4		3				
						12		4		3				
						8		4		2				
G-3,G-14 (G-20 WOULD NOT RUN IN COMBINATION WITH EITHER 3 OR 14).	0.000000	040002		.007376		.007666	258498	.090629	.060237	.010374		.145591		.017240
				4		5	151	19	57	9		6		7
				2		2	7	2	5	2		4		5
				2		2	5	2	4	2		4		5
						2	5	2	4	2		4		5
						2	4	2	4	2		3		4
						2	4	2	4	2		3		4
						2	4	2	3	2		3		4
G-3, G-4	0.049946	040003		.000086		.005116		.066901		.001556		.344753		.016069
G-3, G-20	0.089160					3		4		2		24		9
G-14, G-20	0.299961					3		1		1		4		2
G-3, G-14 G-20	0.467983					3		1		1		4		2
						3		1		1		4		2
						3		1		0		4		2
						3		1		0		4		2
						3		1		0		3		2

FIGURE 6. TABULATION OF VARIATE DIFFERENCES. G-2, G-12, & G-19 ARE THE ASKANIAS. G-3, G-14, & G-20 ARE THE CONTRAVES.

ESTIMATING THE PARAMETERS OF A MODIFIED POISSON DISTRIBUTION*

A. Clifford Cohen, Jr.
University of Georgia

Errors in observing and reporting sample data often complicate the problem of estimating parameters of the distribution being sampled. If neglected, such errors may lead to seriously biased estimates. There exists a large general class of such estimation problems involving numerous different distributions, different types and varying degrees of observational errors. This paper is limited, however, to maximum likelihood estimation in a Poisson distribution which has been modified to the extent that a proportion θ of the ones are reported as being zeros. An inspector who sometimes fails to see or at least fails to report items containing only a single Poisson distributed defect, while correctly observing and reporting results of inspecting items containing two or more defects, produces sample data of the type under consideration. Estimators are derived both for the Poisson parameter and for θ. Asymptotic variances and covariances are derived and an illustrative example is included.

1. INTRODUCTION. In observing a Poisson distributed random variable, it sometimes happens that values of one are erroneously observed or at least reported as being zeros. For example, in determining the number of defects per unit or item examined, an inspector may err by reporting units which actually contain a single defect as being perfect or free of defects. Of course there is also a similar possibility of erroneous observation when the actual number of defects per unit is in excess of one, but here we are concerned only with the case in which some though not necessarily all ones are reported as zeros.

Suppose the number of defects actually present per unit is a Poisson distributed random variable with parameter λ, and that the probability of misclassifying an item containing one defect by reporting it as containing zero defects is θ. The probability function of the random variable x, the observed (reported) number of defects per item, may then be written as

$$(1) \qquad p(x;\lambda,\theta) = \begin{cases} e^{-\lambda}(1 + \theta\lambda), & x = 0, \\ (1 - \theta)\lambda e^{-\lambda}, & x = 1, \\ e^{-\lambda}\lambda^x/x!, & x = 2,3,\ldots, \end{cases}$$

where $\lambda > 0$ and $0 \leq \theta \leq 1$.

In an abstract sense, (1) may simply be considered as the probability function of a two parameter modified Poisson distribution, and in this

---

* This paper was originally published in the Journal of the American Statistical Association (March 1960). Permission to reproduce it here is greatly appreciated by the editors.

paper we are concerned with maximum likelihood estimation of its two parameters λ and θ. The problem under consideration here is a special case of a more general class of estimation problems involving erroneous sample observation which has been encountered, for example, by Neyman and Scott [6] in connection with counting galaxy images on photographic plates and by Toulouse [9] in connection with attribute sampling. It is closely related to the estimation of the Poisson parameter from truncated and censored samples, a problem which received attention from David and Johnson [4], Moore [5], Plackett [7], Rider [8], this writer [2], [3], and various others.

2. DERIVATION OF ESTIMATORS. Consider a sample consisting of N observations of the random variable x with probability function (1) in which n_0 designates the number of zero observations and n_1 the number of ones. The likelihood function for such a sample is

$$P(x_1, \ldots x_N; \lambda, \theta) = \left[e^{-\lambda}(1 + \theta\lambda)\right]^{n_0} \left[(1 - \theta)\lambda e^{-\lambda}\right]^{n_1} \Pi^* e^{-\lambda} \left(\lambda^{x_i}\right) / x_i!,$$

where Π^* is the product over all x's that are neither 0 nor 1. We write this result in simpler form as

$$P(x_1, \ldots x_N; \lambda, \theta) = e^{-N\lambda}(1 + \theta\lambda)^{n_0}(1 - \theta)^{n_1} \lambda^{\sum_1^N x_i} \left[\Pi^* x_i!\right]^{-1}. \tag{2}$$

Taking logarithms of (2), differentiating with respect to λ and θ in turn, and equating to zero yields the estimating equations

$$\begin{aligned} \partial L/\partial\lambda &= -N + n_0\theta/(1 + \theta\lambda) + \sum_1^N x_i/\lambda = 0, \\ \partial L/\partial\theta &= n_0\lambda(1 + \theta\lambda) - n_1/(1 - \theta) = 0 \end{aligned} \tag{3}$$

where L is written for ln P.

The required M.L. estimators $\hat{\lambda}$ and $\hat{\theta}$, when they exist, will be found by simultaneously solving these two equations. We follow the customary notation of employing $(\hat{\ })$ in this paper to distinguish maximum likelihood estimators from the parameters estimated.

To facilitate their solution, the above equations are reduced to

$$\begin{aligned} \lambda^2 - (\bar{x} - 1 + n_0/N)\lambda - (\bar{x} - n_1/N) &= 0, \\ \theta = \left[n_0 - n_1/\lambda\right]/(n_0 + n_1), \end{aligned} \tag{4}$$

where $\bar{x}$ is the sample mean ($\bar{x} = \Sigma_1^N x_i/N$).

The first equation of (4) results from eliminating θ between the two equations of (3), while the second results from solving the second equation of (3) for θ. A similar pair of equations can be obtained by first eliminating λ between the two equations of (3) and thus obtaining an equation which is quadratic in θ. Estimates are easier to calculate, however, using the results given above in (4).

We note that $(\bar{x}-1+n_0/N)>0$ and $(\bar{x}-n_1/N)>0$ except when (i) all sample observations are zeros, or (ii) all observations are ones. With these two exceptions, the coefficients of the first equation of (4), which is quadratic of the form $g(\lambda) = 0$, thus exhibit one change of sign, and likewise the coefficients of $g(-\lambda)$ exhibit one change of sign. It then follows from Descartes' well known "rule of signs" that $g(\lambda) = 0$ has exactly one positive and one negative root. The positive root of this equation is the required estimator of λ, and on solving by means of the quadratic formula, we obtain

$$\hat{\lambda} = \left[(\bar{x} - 1 + n_0/N) + \sqrt{(\bar{x} - 1 + n_0/N)^2 + 4(\bar{x} - n_1/N)}\right]/2. \tag{5}$$

With $\hat{\lambda}$ thus determined, the second equation of (4) enables us to calculate

$$\hat{\theta} = (n_0 - n_1/\hat{\lambda})/(n_0 + n_1). \tag{6}$$

When $\theta = 0$, (1) becomes the ordinary Poisson probability function without modification, in which case the first equation of (3) yields the familiar estimator $\hat{\lambda} = \bar{x}$. We now turn our attention to three special types of samples, two of which were listed as exceptions in the preceding paragraph. Although samples of these types are unlikely to arise in practical applications envisioned for the results of this paper, they are of theoretical interest and are considered here for that reason.

Special type (i). All observations are zeros; $n_1=0$, $n_0=N$, and $\bar{x}=0$. The likelihood equation (2) for a sample of this type becomes

$$P = e^{-N\lambda}(1 + \theta\lambda)^N.$$

On taking logarithms, differentiating with respect to λ and θ in turn and equating to zero, estimating equations corresponding to (3) become

$$-N + N\theta/(1 + \theta\lambda) = 0,$$

$$N\lambda(1 + \theta\lambda) = 0.$$

Maximum likelihood estimates $\hat{\lambda}$ and $\hat{\theta}$ do not exist in this case, however, since the above estimating equations are simultaneously satisfied only when $\lambda = 0$ and $\theta = 1$, whereas $p(x;\lambda,\theta)$ is defined only for $\lambda>0$.

Special type (ii). All observations are ones; $n_0=0$, $n_1=N$, and $\bar{x}=1$. Maximum likelihood estimates $\hat{\lambda}$ and $\hat{\theta}$ fail to exist in this case also since estimating equations (3) are not simultaneously satisfied by any pair of values of λ and θ for which $p(x;\lambda,\theta)$ is defined. Although the first equation of (3) with $n_0 = 0$ is satisfied when $\lambda = 1$, the second is only satisfied in the limit as $\theta\to\infty$, whereas $p(x;\lambda,\theta)$ is defined only for $0 \le \theta \le 1$.

Special type (iii). No zeros or ones are observed; $n_0=n_1=0$. In this case the likelihood equation (2) is independent of θ, which therefore cannot be estimated from available sample information. The Poisson parameter, however, is estimated by (5), which for a sample of this type, reduces to $\hat{\lambda} = \bar{x}$.

It is not difficult to construct other samples for which (5) and (6) fail to give acceptable estimates of λ and θ. However, when N is large such samples will be very improbable and their occurrence in practical applications should be interpreted as a suggestion that probability function (1) might not be applicable to the random variable actually observed.

3. SAMPLING ERRORS OF ESTIMATES. The asymptotic variance-covariance matrix of $(\hat{\lambda},\hat{\theta})$ is obtained by inverting the information matrix whose elements are negatives of expected values of the second order derivatives of logarithms of the likelihood function.

The second partial derivatives of L follow from (3) as

$$\partial^2 L/\partial\lambda^2 = -n_0\theta^2/(1+\theta\lambda)^2 - N\bar{x}/\lambda^2,$$

$$\partial^2 L/\partial\theta^2 = -n_0\lambda^2/(1+\theta\lambda)^2 - n_1/(1-\theta)^2, \tag{7}$$

$$\partial^2 L/\partial\lambda\partial\theta = \partial^2 L/\partial\theta\partial\lambda = n_0/(1+\theta\lambda)^2.$$

Since $E(\bar{x}) = \lambda(1-\theta e^{-\lambda})$, $E(n_0) = Ne^{-\lambda}(1+\theta\lambda)$, and $E(n_1) = N(1-\theta)\lambda e^{-\lambda}$, where E( ) denotes expected value, elements of the information matrix follow from (7) as

$$E(-\partial^2 L/\partial\lambda^2)/N = (1+\theta\lambda-\theta e^{-\lambda})/\lambda(1+\theta\lambda),$$

$$E(-\partial^2 L/\partial\theta^2)/N = \lambda e^{-\lambda}(1+\lambda)/(1+\theta\lambda)(1-\theta), \tag{8}$$

$$E(-\partial^2 L/\partial\lambda\partial\theta)/N = E(-\partial^2 L/\partial\theta\partial\lambda)/N = -e^{-\lambda}/(1+\theta\lambda).$$

On inverting the information matrix, the asymptotic variances and covariance follow as

$$V(\hat{\lambda}) \sim \lambda(1+\lambda)/N(1+\lambda-e^{-\lambda}),$$

$$V(\hat{\theta}) \sim (1+\theta\lambda-\theta e^{-\lambda})(1-\theta)/N\lambda e^{-\lambda}(1+\lambda-e^{-\lambda}), \tag{9}$$

$$\mathrm{Cov}(\hat{\lambda},\hat{\theta}) \sim (1-\theta)/N(1+\lambda-e^{-\lambda}).$$

The correlation coefficient between estimates $\hat{\lambda}$ and $\hat{\theta}$ follows as

$$\rho_{\hat{\lambda},\hat{\theta}} = \mathrm{Cov}(\hat{\lambda},\hat{\theta})/\sqrt{V(\hat{\lambda})V(\hat{\theta})} \sim \sqrt{(1-\theta)e^{-\lambda}/(1-\lambda)(1+\theta\lambda-\theta e^{-\lambda})}. \tag{10}$$

The variances and covariance given in (9) and the correlation coefficient given in (10) are applicable in all cases where maximum likelihood estimators $\hat{\lambda}$ and $\hat{\theta}$ exist. Even with samples of special type (iii), $V(\hat{\lambda})$ as given in (9) is applicable. Since N, the total sample size, is fixed n_0 and n_1 are random variables and although they may assume the value zero in particular samples, their expected values as given in the preceding paragraph are in excess of zero. Of course $E(n_0)\to 0$ and $E(n_1)\to 0$ as $\lambda\to\infty$.

Furthermore, when λ is large $V(\hat{\lambda})$ as given by (9) differs but slightly from λ/N which applies when λ is estimated from a sample of size N from an ordinary Poisson distribution without modification.

4. AN ILLUSTRATIVE EXAMPLE. To illustrate the practical application of results of this paper, data from Bortkiewicz's [1] classical exmaple on deaths from the kick of a horse in the Prussian Army have been suitably altered. The origianl data were collected from records of a certain group of ten Prussian Army Corps over the twenty year period 1875-1894. The study thus included 200 annual reports; that is, 200 observations of the random variable involved. For the purpose of this illustration it has been assumed that twenty of the records which should have shown one death each were in error by reporting no deaths. Both the original and the altered data for this example are given below.

Number Deaths per Army Corps per Year	Number Observations	
	Original Data	Altered Data
0	109	129
1	65	45
2	22	22
3	3	3
4	1	1
5	0	0

Summarizing the altered (misclassified) data, we have: $n_0 = 129$, $n_1 = 45$, $N = 200$, $\bar{x} = 102/200 = 0.51$, $n_0/N = 0.645$, $n_1/N = 0.225$, $(\bar{x} - 1 + n_0/N) = 0.155$, and $(\bar{x} - n_1/N) = 0.285$. On substituting these values into (5), we calculate

$$\hat{\lambda} = \left[0.155 + \sqrt{0.155^2 + 4(0.285)}\right]/2 = 0.617$$

Subsequent substitution into (6) yields

$$\hat{\theta} = (129 - 45/0.617)/(129 + 45) = 0.322.$$

The estimate $\hat{\lambda} = 0.617$, obtained above is to be compared with 0.610 which follows from the original unaltered data. The estimate $\hat{\theta} = 0.322$ is to be compared with $20/65 = 0.308$, which is the proportion of ones that were misclassified in the process of altering the original data for this illustration.

With λ and θ replaced by their estimates $\hat{\lambda}$ and $\hat{\theta}$, (9) and (10) enable us to calculate

$$V(\hat{\lambda}) \doteq 0.0046,$$
$$V(\hat{\theta}) \doteq 0.0097,$$
$$\mathrm{Cov}(\hat{\lambda},\hat{\theta}) \doteq 0.0031,$$
$$\rho_{\hat{\lambda},\hat{\theta}} \doteq 0.47.$$

$V(\lambda) \doteq 0.0046$ as calculated above for $\hat{\lambda}$ based on the altered data is to be compared with $V(\hat{\lambda}) \sim \lambda/N \doteq 0.610/200 = 0.00305$ for $\hat{\lambda}$ based on the complete (unaltered) sample.

REFERENCES

1 Bortkiewicz, L. von, Das Gesetz der Kleinen Zahlen, Leipzig: Teubner, 1898.

2 Cohen, A. C., Jr., "Estimation of the Poisson parameter from truncated samples and from censored samples," Journal of the American Statistical Association, 49)1954), 158-68.

3 Cohen, A. C., Jr., "Estimating the parameter in a conditional Poisson distribution," In press for publication in Biometrics.

4 David, F. N. and Johnson, N. L., "The truncated Poisson," Biometrics, 8 (1952), 275-85.

5 Moore, P. G., "The estimation of the Poisson parameter from a truncated distribution," Biometrika, 39 (1952), 247-51.

6 Neyman, Jerzy and Scott, Elizabeth L., "Large scale organization of the distribution of galaxies," Handbuch der Physik, 53 (1959), 416-44.

7 Plackett, R. L., "The truncated Poisson distribution," Biometrics, 9 (1953), 485-88.

8 Rider, Paul R., "Truncated Poisson distributions," Journal of the American Statistical Association, 48 (1953), 826-30.

9 Toulouse, Julian H., "Psychological bias in attribute sampling," Industrial Quality Control, 14 (June 1958), 1-8.

THE DETECTION OF GUESS RESPONSES IN THE RATING OF STATEMENTS BY THE METHOD OF SUCCESSIVE CATEGORIES

Lee E. Paul and Howard W. Hembree
QM Research and Engineering Field Evaluation Agency

This problem arose in the process of constructing equal interval rating scales for the clothing characteristics of fit, comfort, protection, durability and over-all acceptability. Guilford's method of successive categories was used to determine the scale values of a number of descriptive statements covering the full range of the dimensions studied. To accomplish this, the statements were administered to a sample similar to the population which will eventually use the scales and the subjects were asked to place each statement in one of 11 categories, category 1 the least favorable, category 11 the most favorable. A numerical value was computed for each of the 11 categories such that they tended to normalize the frequency distribution of the ratings for each statement. These cagegory values were used to determine the mean and standard deviation of each statement. To construct a scale of any length, one simply selects statements such that the means are equidistant and the standard deviations are low.

Past experience in the administration of questionnaires to enlisted personnel reveals some small proportion who are not highly motivated in pursuing a task, the goal of which seems rather remote. Some of these men might be conservatively described as indifferent. However, since they have been told to rate these statements they must comply one way or another. It then becomes necessary to identify those respondents who did not rate the statements according to the instructions, either because they did not understand the task the English language or because of a lack of motivation. There are two ad hoc methods for detecting these non-conformists. One is simply to look at their responses, (see Table 1 at end of this article), with an eye to detecting patterns, that is, some mechanical scheme for responding that the subject feels will go undetected before he leaves the test session. Subject number 25 chose a rather unimaginative method. To cope with the more complex "guessers" another method has been used which consists of taking pairs of statements, one of each pair obviously favorable, another obviously unfavorable and looking for reversals in the ratings. With a large number of respondents, this can immediately be seen to be rather tedious and time consuming. The most important shortcoming of these methods, however, is that they are not objective in that there is no standardized procedure for their application nor do they provide any information on a cutting point, i.e., some score that indicates the respondent doesn't belong in a normal conscientious population.

The problem, then, is to identify two populations, one which understands the instructions, the English language and is motivated enough to make an honest effort, the other having shortcomings on one or more of these characteristics.

The solution proposed here is quite simple. A number of criterion statements were selected such that 50% or more of the total sample rated each statement in the two most extreme categories: 1 and 2, or 10 and 11,

(see Table 2). This is, of course, somewhat arbitrary. The method further defines a "guess" as a rating of 6 or less on a favorable criterion statement, 6 or more on an unfavorable one.

The criterion statements shown in Table 1 are 1, 5, 7, 10, 12, 15, 16, 22, only 8 meeting the requirements in this case. Now one simply descends the criterion columns and circles any response of 6 or less or 6 or more, depending on whether the majority finds statements are favorable or unfavorable (6 was included in both to catch the guesser whose system, if he has one, includes a lot of 6's). Simply counting the circles in a row provides a guess score for each subject.

It was noted earlier that each respondent rated two sets of statements. The rating of statements was carried out in 5 sessions as shown in Table 3.

Table 3

Session	Characteristics	Criterion Statements	R	N
1	Fit	10	.78	56
	Over-all Acceptability	8		
2	Durability	10	.929	61
	Protection	9		
3	Comfort	10	.668	51
	Protection	9		
4	Fit	10	.865	67
	Acceptability	8		
5	Durability	10	.534	59
	Comfort	10		
				294

Ten of the "fit" statements met the conditions for criterion statements, 8 for "over-all acceptability," 10 for durability," 9 for "protection" and 10 for "comfort." N is the number of subjects in each session, while R is the product-moment correlation of the guess scores obtained from the two sets of criterion statements. It seems apparent that the respondents were fairly consistent throughout the rating and that the criterion statements from the different characteristics were measuring the same thing with the possible exception of those from "comfort."

Figure 1 represents the frequency distribution of guess scores for all five characteristics with each individual represented twice, once for each characteristic rated. This distribution is a bimodal J curve with the primary mode at 0 and a smaller one at 4. Each man rated an average of 9.39 criterion statements and had 6 chances in 11 of getting "caught" at each guess. Thus, the mean for all guessers should be 9.39 x 6/11 or 5.11.

In order to get better separation between the guessers at the rest of the population, the guesses of each individual on the two characteristics he rated were combined and the frequency distribution showed in Figure 2. The two populations already suggested now seem evident. One population, the larger, has a J shaped distribution, which brings to mind Allport's J curve of social conformity. One explanation of this distribution is that it is a normal distribution that is quite insensitive at one end. In this case it seems probable that not all those with 0 scores were equally able or conscientious.

The other distribution, which consists of "guessers," is apparently normal. This distribution is based on an average of 18.78 criterion statements for each subject and should have a mean of 10.22 or 18.78 x 6/11. A binomial expansion indicates the standard deviation of this distribution should be approximately 2.2. $\left[\sqrt{NPQ} = (18.78 \times 6/11 \times 5/11)^{\frac{1}{2}}\right]$ Actually the S.D. will be slightly larger since some subjects were scored on 18 criterion statements, some 19, and some 20.

The dotted line in Figure 2 represents a smoothed extrapolation of the J curve. The inset distribution of guessers is the total curve less the smoothed J curve. While the inset curve doubtlessly includes some misclassified subjects, it is a reasonably close approximation of a symmetrical normal curve with a mean of about 10 and a standard deviation of a little over two.

Using a cutting point of 6 (mean - S.D. = 10.22 - 1.92 x 2.2) ought to eliminate about 97% of the guessers. This figure is very close to the intersection of the two curves and so approximates a maximum likelihood ratio.

As a somewhat independent evaluation of the guess scores, some product-moment correlation coefficients were computed between the median rating of the 32 statements for "fit" and the ratings of a number of individuals. These respondents were selected so as to include different guess scores. The correlations were plotted against the guess scores, and the results are shown in Figure 3. While it is obvious that people with high guess scores do not agree with the majority as expressed by median ratings, the number of near 0 and negative correlations reveal that many respondents actually guessed, in the literal sense, throughout the task.

The application of this procedure led to the elimination of the ratings of approximately 15% of the original sample. Since the ratings of most of the rejectees were apparently random with respect to the median value of the statements, it is believed that their inclusion in further computations would have led to a spuriously high estimate of the standard deviations of the statements.

STATEMENTS (O.A.I) ⟶ (Table 1)

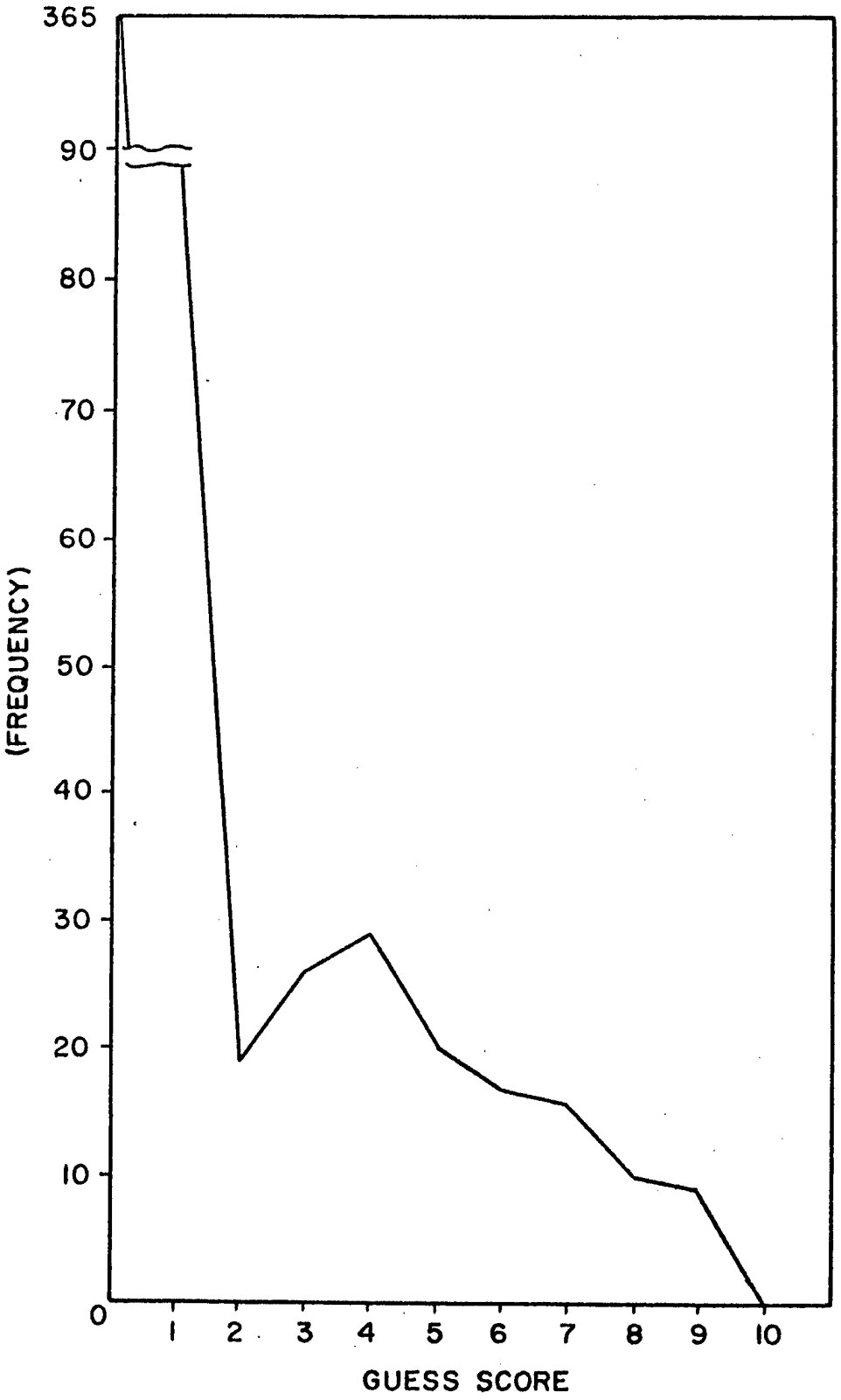
SUBJECTS ↓

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
1	9	4	8	7	(8)	7	5	9	9	10	5	3	6	8	11	2	4	11	5	9	4	3	4	6	4	7	6	3	4	7	7	6
2	(4)	11	6	4	5	7	5	6	4	7	4	4	5	6	(4)	(8)	9	8	10	8	11	(11)	8	10	9	10	8	8	9	8	10	11
3	8	6	5	5	(6)	8	(6)	5	5	(6)	4	(7)	7	7	(4)	(7)	5	4	4	5	5	(6)	7	8	5	9	11	8	6	4	7	4
4	11	4	9	5	1	5	1	11	4	11	5	1	5	7	10	2	4	6	1	5	5	2	4	6	5	5	8	4	5	5	9	4
5	(6)	4	2	5	(7)	6	1	6	6	7	5	3	3	7	(3)	6	7	6	2	3	2	2	2	7	6	6	6	6	8	8	6	1
6	9	6	7	4	5	6	1	7	6	8	4	1	4	7	11	3	5	5	4	5	4	1	4	4	2	4	5	6	6	6	7	6
7	11	6	10	3	1	6	2	8	5	11	4	1	6	10	11	4	3	5	5	6	2	2	4	6	2	7	7	4	4	4	8	5
8	11	9	8	2	1	7	2	9	4	10	3	1	6	8	10	4	4	8	4	6	2	2	4	9	2	8	7	3	6	4	7	5
9	11	5	9	2	1	5	1	7	4	11	5	1	6	7	11	1	1	9	2	7	4	1	2	5	5	10	9	2	3	3	6	4
10	8	7	11	6	5	9	(6)	8	5	(6)	8	(8)	9	10	7	(11)	6	5	8	7	5	(7)	7	7	5	4	4	5	8	9	7	9
11	11	4	8	6	1	7	3	8	4	8	2	1	5	9	11	1	4	5	1	5	1	1	3	9	4	9	7	4	3	2	7	1
12	8	6	7	6	(7)	6	(6)	9	6	8	5	(6)	9	7	10	(6)	7	9	7	5	4	3	7	5	5	9	10	6	4	8	4	5
13	(4)	10	9	5	1	7	1	10	4	11	4	1	6	8	11	1	7	8	4	5	5	4	4	6	4	8	8	4	7	4	7	6
14	11	2	8	3	1	6	1	9	2	11	2	1	7	8	11	1	2	9	4	9	3	2	3	7	4	8	7	4	6	4	5	3
15	11	7	9	7	1	6	1	10	7	10	1	1	7	9	11	1	2	8	5	7	2	2	5	10	4	11	10	7	2	3	9	6
16	11	4	10	4	1	5	1	9	5	11	4	1	8	8	11	1	3	4	1	3	4	1	8	6	3	6	8	3	5	4	10	4
17	8	5	10	2	1	8	1	6	5	10	3	2	6	8	11	4	5	6	5	4	3	2	3	7	4	9	8	3	8	4	6	3
18	9	5	6	3	1	6	2	8	5	10	3	1	8	5	10	4	4	7	2	6	3	2	3	6	4	8	9	6	5	4	6	6
19	11	6	10	2	(11)	6	1	10	2	10	2	1	8	6	11	1	3	6	3	6	2	1	3	11	3	10	6	2	3	3	6	3
20	11	9	10	5	2	7	2	10	4	10	4	1	8	9	10	2	5	7	3	6	5	2	4	6	5	7	8	4	8	3	9	6
21	11	6	10	3	1	6	1	10	4	10	2	1	6	9	11	2	2	10	2	4	2	1	3	6	3	9	10	1	4	3	6	2
22	9	5	8	4	1	7	3	8	5	8	3	2	8	7	10	3	6	4	7	9	3	(7)	6	10	6	6	5	2	7	2	6	2
23	11	10	9	3	2	6	1	10	4	10	3	2	6	9	11	2	3	10	4	5	4	3	4	6	4	9	8	3	9	5	9	8
24	10	X	11	5	1	4	2	2	4	11	5	1	6	9	10	2	4	6	2	5	2	2	4	3	6	5	9	4	4	3	9	5
25	(1)	1	1	5	1	1	1	1	1	(1)	1	1	1	1	(1)	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	5
26	8	1	9	5	1	7	2	10	5	10	4	1	6	8	10	1	1	4	1	5	1	1	3	6	6	8	8	7	4	3	8	6
27	11	7	10	3	2	8	2	10	5	11	3	1	7	8	10	3	3	7	3	6	4	1	5	8	2	10	8	4	6	3	10	2
28	11	3	10	8	1	6	2	10	7	11	3	1	6	9	11	2	5	10	4	7	4	2	4	7	4	9	10	7	8	4	5	7
29	11	4	9	3	1	6	1	10	4	10	3	1	6	8	11	3	3	8	1	6	2	1	3	6	3	8	8	3	5	3	8	4
30	9	4	8	3	1	4	1	7	4	10	3	1	5	5	10	2	3	7	2	6	3	1	4	6	3	7	8	3	5	2	5	3
31	11	4	11	2	2	6	2	9	4	11	6	2	8	9	11	1	2	3	2	8	2	1	2	8	4	8	9	4	5	5	10	5
32	11	7	10	5	1	7	(10)	9	4	10	5	1	6	9	11	3	4	7	2	5	4	2	4	6	4	9	8	4	4	4	9	5
33	11	5	11	2	1	6	1	10	3	9	1	(6)	2	1	(4)	(6)	4	8	8	4	4	4	6	4	4	11	5	1	4	8	11	1
34	(6)	7	7	4	5	8	4	9	7	9	5	4	9	10	10	(11)	8	7	8	3	6	3	5	9	7	8	10	8	7	7	11	10
35	(6)	5	4	5	(6)	4	5	3	4	5	4	4	3	3	(4)	4	1	3	5	6	5	4	4	3	1	2	4	3	3	4	4	5
36	11	9	10	3	1	7	1	10	3	10	6	1	8	9	11	2	3	8	4	5	2	3	4	6	4	9	9	4	4	3	9	8
37	(1)	5	1	3	1	1	1	3	3	(1)	4	4	1	1	(1)	1	1	1	6	1	3	1	1	1	1	1	1	2	1	8	1	7
38	11	7	9	2	1	5	1	9	4	8	3	1	6	8	10	1	1	6	1	7	4	2	4	6	4	8	8	4	9	4	7	4
MEDIAN	10.8	5.5	8.9	3.8	1.4	6.0	1.8	8.9	4.2	9.8	3.8	1.3	6.1	7.8	10.6	2.3	3.8	7.1	3.6	5.6	3.5	2.2	3.7	6.0	3.5	7.6	7.8	3.4	5.3	3.7	7.1	4.6

(Table 2)

Statements	Percent of Ratings in Each Category										
	1	2	3	4	5	6	7	8	9	10	11
*PERFECT IN EVERY RESPECT	0.8	1.6	0.8	2.5	0.8	5.7	1.6	4.9	6.6	7.4	67.2
NOT GOOD ENOUGH FOR EXTREME CONDITIONS	1.6	7.4	7.4	18.0	16.4	15.6	13.1	4.1	8.2	4.1	4.1
VERY GOOD	0.8	0.8	2.5	1.6	3.3	4.1	7.4	20.5	24.6	24.6	9.8
BARELY ADEQUATE	4.1	11.4	27.1	23.8	23.0	5.7	1.6	0.8	1.6	0.0	0.8
*VERY UNSATISFACTORY	54.4	21.1	0.8	0.8	5.7	4.9	2.5	3.2	0.0	3.2	3.2
MODERATELY GOOD	2.5	0.8	0.8	8.2	22.1	30.3	21.3	9.0	3.3	1.6	0.0
*VERY POOR	41.0	28.7	11.5	3.3	4.1	2.5	3.3	1.6	0.0	2.5	1.6
UNUSUALLY GOOD	0.0	2.5	4.1	0.0	5.8	8.3	9.1	12.4	23.1	28.9	5.8
NOT QUITE ADEQUATE	4.1	5.7	11.5	43.4	19.7	7.4	4.9	2.5	0.8	0.0	0.0
*EXTREMELY GOOD	0.8	1.7	0.8	0.0	3.3	4.1	7.4	14.1	9.1	29.8	28.9
NOT VERY SATISFACTORY	3.3	10.7	26.4	34.7	13.2	5.8	2.5	2.5	0.0	0.8	0.0
*EXTREMELY POOR	60.7	18.9	2.5	4.9	1.6	4.1	0.8	2.5	1.6	2.5	0.0
ABOUT AVERAGE	0.8	2.5	3.3	4.9	16.4	36.1	11.5	14.8	4.9	2.5	2.5
VERY GOOD IN MOST RESPECTS	1.7	1.7	0.8	0.8	5.8	11.6	20.7	24.0	19.8	9.9	24.8
*EXCELLENT	0.8	0.8	1.6	3.3	1.6	2.5	1.6	2.5	6.6	21.3	57.4
*BETTER THAN NOTHING	27.1	28.7	14.8	10.7	4.9	5.7	1.6	2.5	2.5	0.0	1.6
NEEDS MAJOR CHANGES	10.7	9.9	22.3	24.8	11.6	5.0	5.8	2.5	3.3	1.7	2.5
MORE THAN ADEQUATE	1.7	0.0	3.3	9.1	11.6	15.7	17.4	19.8	14.9	4.1	2.5
NOT GOOD ENOUGH FOR GENERAL USE	16.4	13.1	18.9	23.0	9.0	4.9	5.7	4.1	2.5	1.6	0.8
ADEQUATE	1.6	0.0	6.5	10.7	28.7	30.3	9.0	4.9	4.9	2.4	0.8
BARELY ACCEPTABLE	5.7	24.6	19.7	22.1	18.0	4.1	1.6	1.6	0.8	0.0	1.6
*POOR	28.7	29.5	20.5	5.7	3.2	2.5	3.3	2.5	0.0	0.8	3.3

*Criterion statements



FREQUENCY DISTRIBUTION OF GUESS SCORES FOR ALL FIVE CHARACTERISTICS WITH EACH INDIVIDUAL REPRESENTED TWICE, ONCE FOR EACH CHARACTERISTIC RATED. (N=588)

FIGURE 1.

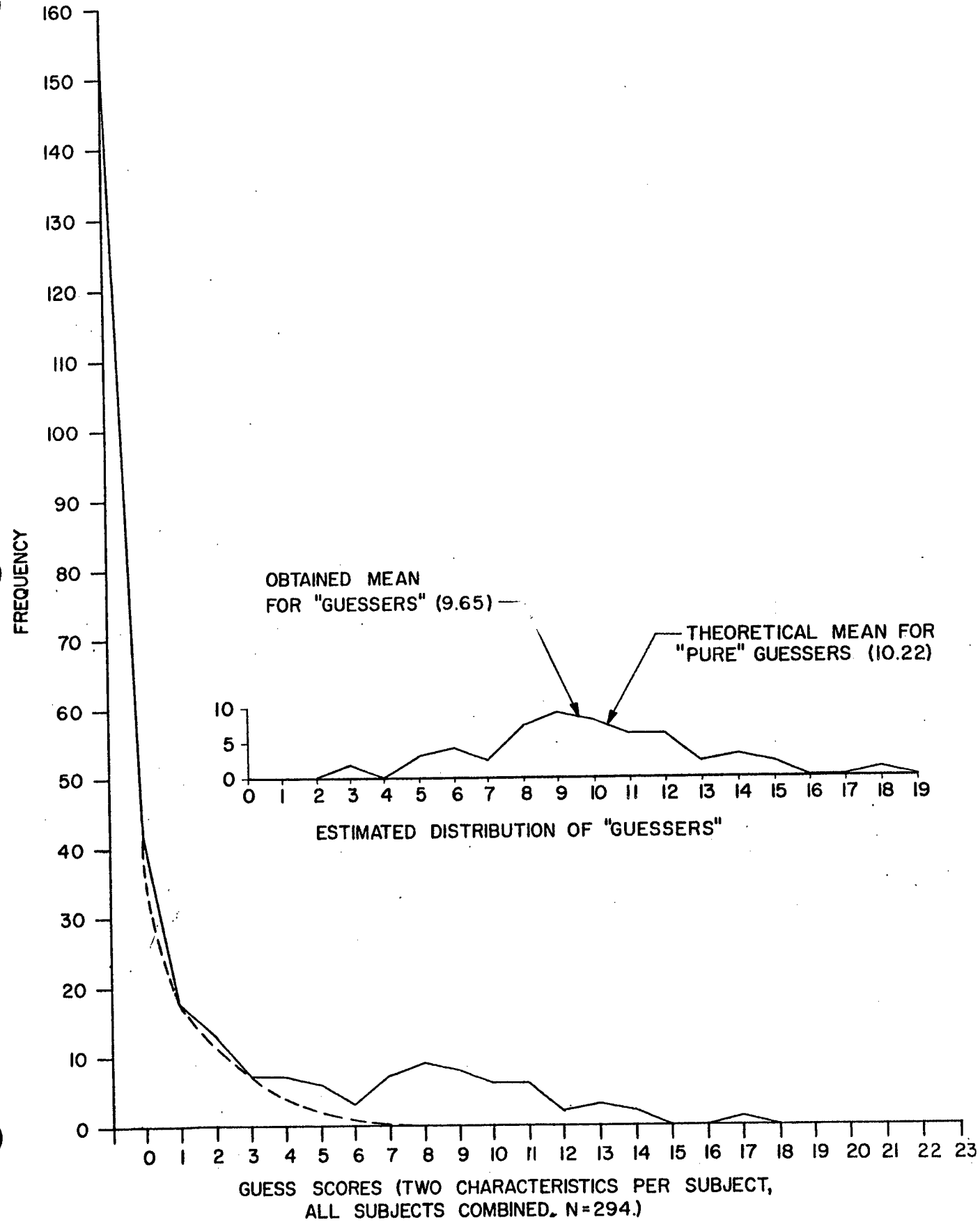
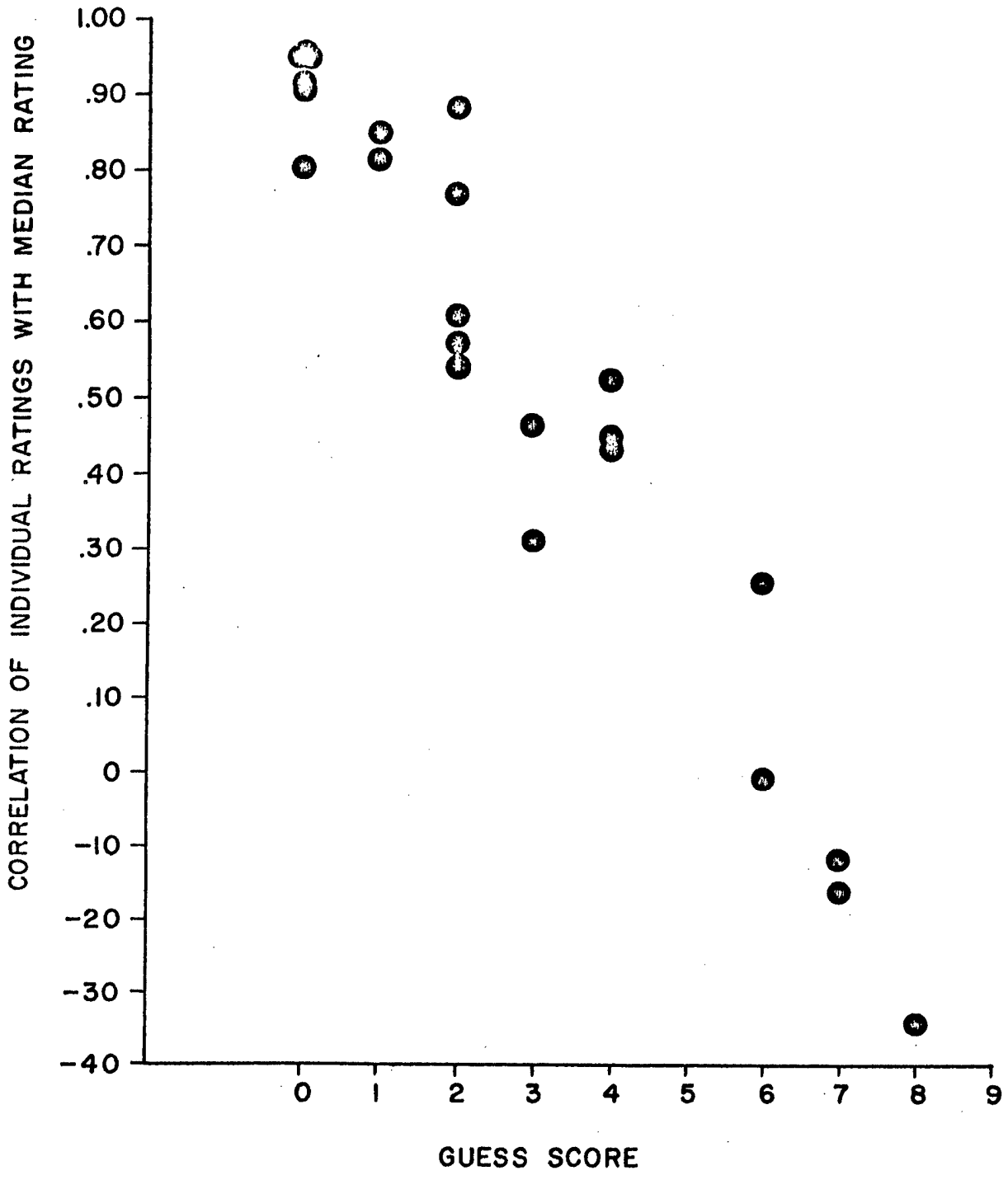


Fig 2. FREQUENCY DISTRIBUTION OF GUESS SCORES



(Figure 3)

CORRELATION OF INDIVIDUAL STATEMENT RATINGS WITH MEDIAN RATINGS PLOTTED AGAINST INDIVIDUAL GUESS SCORES.

DESIGN FOR ESTIMATION BY COVARIANCE TECHNIQUE

M. Rhian
Aerobiology Division, U. S. Army Biological Warfare Laboratories

The problem presented in this paper is similar to those discussed by Cox in the First Conference on Design (1), by Maloney in the Second Conference (2), and by Morrison in Biometrics in 1956 (3). The problem is the estimation of an observation which is purposely not observed. The methods, or mechanics of estimating one or more missing values seems to be of little concern, because there are so many recent descriptions of such procedures. Among these may be mentioned analysis of incomplete data by Wilkinson (4) nature and use of covariance by Cochran (5), and analysis of covariance as a missing plot technique by Coons (6). These and similar articles explore the mathematical bases for calculating missing observations and describe the procedures. These articles also suggested the title of this problem, but your discussion need not be limited to the consideration of covariance analysis.

Specifically, our problem is to estimate the dose of certain microorganisms required to cause disease in certain animals, when members of the animal species cannot be used in direct experimentation. An approach to a solution of this problem may be obtained from brief consideration of the meaning of estimated, values designs used in other situations to obtain doses for estimating, and examples of observed and calculated values.

A few years ago the estimate of a missing value was regarded as a computational convenience and "was not intended as an estimate of the missing datum" (7). This view was challenged in 1954 by Nelder (8) who said that "whether or not x is intended to be an estimate of the missing datum, it is an estimate of the missing datum, and an unbiased one where the mathematical model used is true." This view was supported by Norton (9) who corrected a typographical error in Nelder's formula for calculating the variance of the missing value. Smith (10) then pointed out "that the error variance of the estimate depends on what is is intended to estimate. This must be decided first and then the other aspects fall easily into place." "The variance is, therefore, relevant and Norton's discussion is essentially correct." Smith also pointed out the correct variance test for judging whether an estimated value is preferable to one that was rejected because it seems incompatible with the rest of the data. Rejection of a suspected observation seems to imply that the calculated value is more apt to be valid than the observed value. The view that a calculated value may be "what would have been observed" has not been challenged in Biometrics since 1954. Maybe it will be today.

Cox (1) described two designs for making small samples do double duty, the cross or butterfly design and an x design.

The cross design is illustrated in Slide 1 (at end of this article), taken from Cox's Table 9. He says "assuming no interactions exist it is possible to estimate an expected value for any of the 12 possible combinations of A and B which are given in Table 9 whether or not the combination has data assigned to it." The x design is illustrated in Slide 2, also from Cox (1). In this case three restrictions are necessary to make it

possible to calculate a value for any of the nine combinations. If these designs are valid, they illustrate what may be extreme examples of estimation of about one-half of the total observations.

Morrison describes fractional replications for mixed series and illustrates the estimation of unobserved values (3). One example is taken from Morrison's Table VI. For a 2^4 x 3 experiment, all 48 data points had been obtained then it was decided to set up an estimation design and to compare the results of the half-replication. Slide 3 shows only 1/8 of this table. "The standard deviation of the difference between an observation and an estimate is 0.58. For the 24 points estimated, the maximum discrepancy is equal to about 4 standard deviations (of a difference)." "It may be noted that at the 5% level of significance the results of the half- and full-replicate agree."

To illustrate the anlaysis of incomplete data by covariance Wilkinson (4) used data on blood sugar of rabbits treated with insulin. The original data consisted of observations on 8 rabbits in 4 phases, so there were 32 observations each on percentage fall and initial blood sugar. Three observations on percentage fall in blood sugar were discarded at random, then the corresponding initial values were discarded. The missing values were then estimated by the procedures described by Wilkinson. The values which were observed and discarded, and the values estimated are shown in Slide 4.

These examples from Morrison and Wilkinson indicate that under proper conditions unobserved values can be estimated with acceptable accuracy and precision. We would like to do as well in our situation.

Now what do we have to work with? The basic observations must be dose-response relationships, in which dose is expressed as numbers of deposited organisms and response presumably may be either disease termi-- nated by recovery or disease terminated by death.

In the example chosen for this problem, the doses are numbers of spores deposited and the responses are time to death of the diseased animal. A possible array of data is illustrated in Slide 5. For host species D, doses of organisms are unknown, but the time responses can be obtained from case histories.

As presented in Slide 5 all the missing doses occur in one row, and this seems to complicate the analysis. Can the experiment be designed to change this arrangement?

DeLury has presented procedures for analysis of latin squares when one column or more is missing. Can similar approaches be used in our case?

Slide 1

	B_1	B_2	B_3	B_4
A_1			9 10 12 13 11	
A_2	7 7 8 8 9	9 13 11 11 12	11 12 14 13 16	13 19 15 17 16
			13 14 15 16 19	

The Cross Design (Cox, Table 9).

Slide 2

	B_1	B_2	B_3
A_1	3 8 4 5 7		6 8 13 9 12
A_2		5 11 7 6 8	
A_3	4 7 10 8 5		8 9 13 16 17

X Design (Cox, Table 13)

Slide 3

Observation	Dependent Variable	
	Observed	Estimated
11213	20.15	
11223	28.15	26.36
11212	19.80	20.12
11222	26.70	
11221	19.20	
	25.75	24.40

Comparison of Observed and Estimated Values
(Morrison (3))

Slide 4

Variable	Observed and discarded values	Estimated values
y	33.9	33.2
	24.1	25.6
	35.9	39.8
x	96.9	94.0
	73.9	78.0
	79.9	85.0

Observed and Estimated Values (Wilkinson (4))

Slide 5

Host	Situation A			Situation B		
A	DA_1TA_1	DA_2TA_2	DA_3TA_3	DA_1TA_4	DA_2TA_5	DA_3TA_6
B	DB_1TB	DB_2TB_2	DB_3TB_3	DB_1TB_4	DB_2TB_5	DB_3TB_5
C	DC_1TC_1	DC_2TC_2	DC_3TC_3	DC_1TC_4	DC_2TC_5	DC_3TC_5
D#	XD_1TA_1	XD_2TA_2	XD_3TA_3	XD_1TD_4	XD_2TD_5	XD_3TD_6
N	DN_1TN_1	DN_2TN_2	DN_3TN_3	DN_1TN_4	DN_2TN_5	DN_3TN_6

#Unavailable for direct challenge

Type of data that can be obtained
for estimation of Dose to Host "D".

REFERENCES

1. Cox, Paul C. Some Design Techniques Used for Increasing Cell Size with Special Emphasis in the Missile Field. Proceedings of the First Conference on the Design of Experiments in Army Research, Development, and Testing. Oct. 19-21, 1955. Office of Ordnance Research.

2. Maloney, C. J. Methods of Estimating Lethal Dose for Man. Proceedings of the Second Conference on Design of Experiments in Army Research, Development, and Testing. Oct. 17-19, 1956. Office of Ordnance Research.

3. Morrison, Milton. Fractional Replication for Mixed Series. Biometrics 12, 1-19, 1956.

4. Wilkinson, G. N. The Analysis of Covariance with Incomplete Data. Biometrics 13, 363-72, 1957.

5. Cochran, W. G. Analysis of Covariance: Its Nature and Uses. Biometrics 13, 261-81, 1957.

6. Coons, Irma. The Analysis of Covariance as a Missing Plot Technique. Biometrics 13, 387-405, 1957.

7. Snedecor, G. W. Answer to Query 96. Biometrics 8, 383-4, 1952.

8. Nelder, J. A. A Note on Missing Plot Values. Biometrics 10, 400-01, 1954.

9. Norton, H. W. A Further Note on Missing Data. Biometrics 11, 110, 1955.

10. Smith, H. F. Missing Plot Estimates. Biometrics 13, 115-18, 1957.

11. DeLury, D. B. The Analysis of Latin Squares When Some Observations Are Missing. J. Am. Stat. Assn. 41, 370, 1946.

DESIGN OF AN EXPERIMENT TO EVALUATE A BIO-ASSAY WITH NON-PARALLEL SLOPES

Albert L. Fernelius
Process Research Division, U. S. Army Biological Warfare Laboratories

Graded response virulence estimates of three treatment conditions: ungerminated spores (U), germinated spores (G) and vegetative cells (V) of Bacillus anthracis were made with the mouse as the test animal. For testing virulence the graded response median-time-to-death (MTD) assay takes the form

$$MTD = a \cdot D^{b}$$

where D is the concentration of organisms administered to the host, a is the intercept, and b is the dose-response slope. For a single intraperitoneal test dose consisting of approximately 10^8 cells the MTD values for the three treatment conditions were: U-15.5 hours, G-11.3 hrs, V-8.0 hrs which led to the conclusion that the decreasing order of virulence for treatment condition was $U < G < V$. In subsequent trials, four doses spaced at one log intervals were given to the host. The results in Handout 1 at end of this article indicate that the decreasing order of virulence for a 10^5 dose was $G < U < V$ and the MTD values for this dose were: U - 24.2 hours, G - 42.8 hours, V - 19.8 hours. Plots of the dose-response curves for the three treatment conditions are shown in Handout 2. Estimated slopes of these curves are given in column 5 of Handout 1. Germinated spores generated a response slope approximately two times greater than that given by vegetative cells or ungerminated spores, so it seems possible that any comparison of treatment conditions must be based on sensitivity of the host to changes in dose, i.e., the slope, rather than on MTD values alone. When identical slopes are obtained, then MTD values can be directly compared.

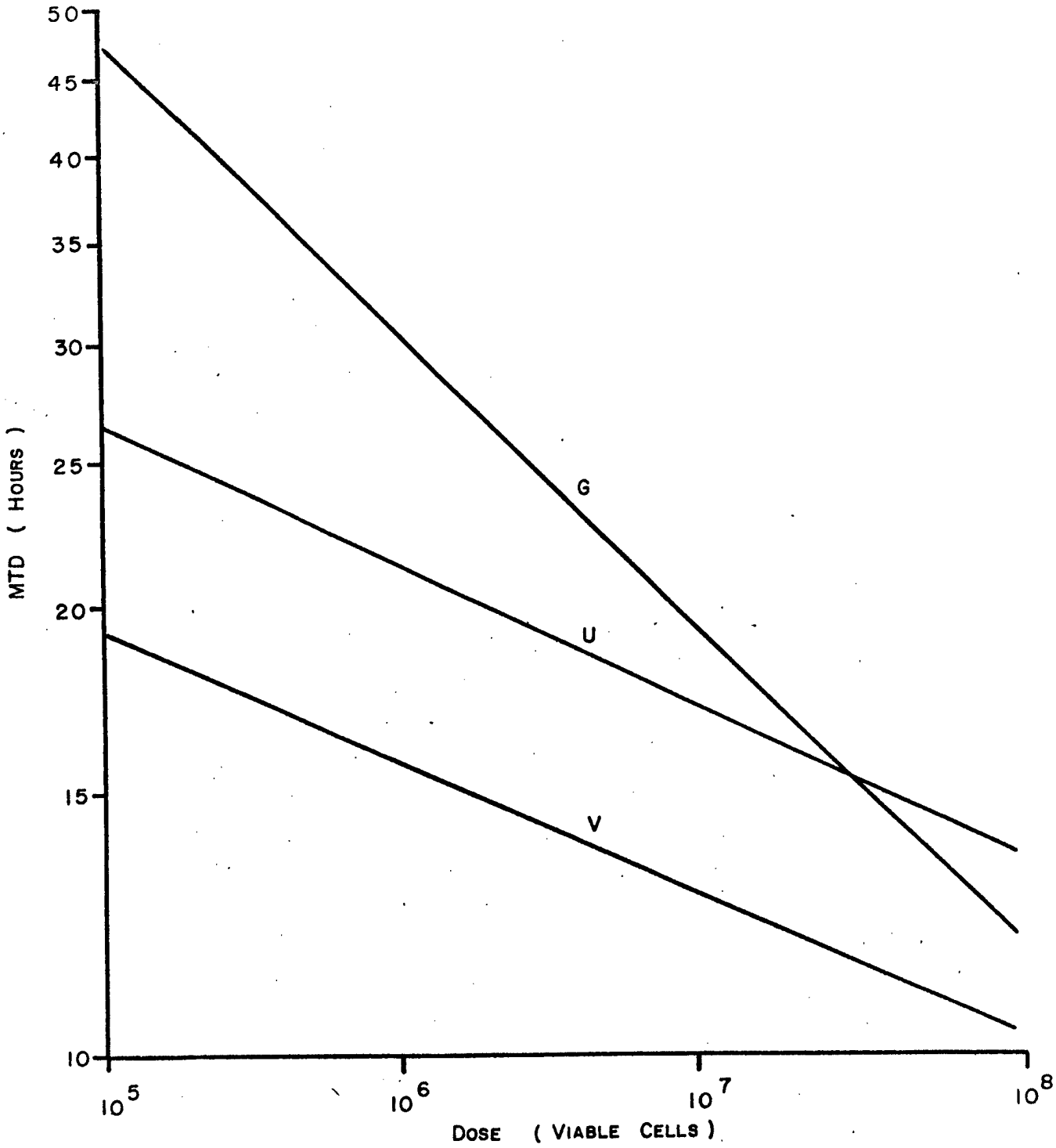
The problem I would like to present to this panel is: How can one design an experiment to compare and evaluate treatment conditions which elicit non-parallel responses in the host? Is there any method of combining the parameters of slope and graded response (MTD) values so that heterogeneous data can be directly comapred? Obviously when MTD values are compared, their relative values will be governed by the point selected on the dose-response curve. How do you determine the dose to titrate, or must one always titrate multiple doses for any comparative purpose?

Handout 1

Mouse Median-Time-To-Death and Log Slope Values for Virulence tests of Bacillus anthracis organisms in three stages of the Spore-Vegetative Cell Cycle

Cyclic stage	$\bar{x}$ Dose* (cells)	$\bar{x}$ MTD (hours)	$\bar{x}$ MTD for a $10^{6.5}$ dose	$\bar{x}$ log slope
Ungerminated	10^8	13.2	19	-.090
	10^7	18.3		
	10^6	22.5		
	10^5	24.2		
Germinated	10^8	11.4	24	-.187
	10^7	22.3		
	10^6	32.6		
	10^5	42.8		
Vegetative	10^8	9.2	14	-.090
	10^7	13.3		
	10^6	17.1		
	10^5	19.8		

* All values are means of three replications.



HANDOUT 2. DOSE RESPONSE FOR UNGERMINATED, GERMINATED, AND VEGETATIVE BACILLUS ANTHRACIS ORGANISMS I.N MICE

THE ORO AIRCRAFT VULNERABILITY EXPERIMENT*

Charles A. Bruce and Bruce Taylor
Operations Research Office, the Johns Hopkins University

INTRODUCTION. This paper describes some techniques planned for the analysis of results from the ORO part of the Aircraft Vulnerability Experiment performed at The Combat Development Experimentation Center.

A major reason for our interest in these techniques is the large amount of data generated during the experiment. The origianl data were taken on 200,000 feet of film and were the equivalent of around two million individual numbers (or readings). After the reduction of this data, which is of necessity a computer operation, there will result about 19,000 numbers. These numbers will be inputs to the techniques described in this paper. We are looking for major trends and highlights in the reduced data. The methods used must both considerably reduce the 19,000 input numbers and also provide valid indications of important conclusions. These conclusions are concerned with tactics and design of Army aircraft, and air defense weapons.

The techniques planned are graphs, curvilinear regression, contingency tests, linear correlation, and analysis of variance. It would simplify matters considerably if the data turns out to be predominantly deterministic and lacking in noticeable change fluctuations. In this case graphs will be drawn for special cases and limiting conditions, and will be followed by a curvilinear regression analysis. However, if sizeable statistical fluctuations appear, contingency tables, linear correlations, and analysis of variance will be used to detect completely random effects and to place some bounds on the fluctuating variables. The problem is not so much one of devising new techniques, but of knowing which of the standard techniques are applicable to this particular experiment.

BACKGROUND INFORMATION. The experiment was undertaken to determine the vulnerability of low flying aircraft to forward area ground fire, and in particular, how this vulnerability depends upon the velocity, altitude, and crossing range of the aircraft with respect to the ground weapons, and also the alert status and line-of-sight terrain masking of the ground troops. The weapons used were Redeye, Quad .50 Cal. machine gun, Twin 40mm antiaircraft weapon, .50 Cal. machine gun on armored personnel carrier, the BAR, and M1 Rifle. The raw data were basically gun-camera film to measure aiming errors, radar and phototheodolite data to give aircraft position, and pen records for timing information. Fig. 1 shows precisely what is meant by mask angle.

In regard to work completed to date, the experiment was designed as a complete factorial with two replications, but only one replication was run due to time and equipment limitations. The experiment was run continuously for a period of one month, during which time approximately 500 single aircraft-passes were made over ground troops. The raw data has been transferred to IBM cards and then to Univac magnetic tape. At the present time, a small amount of data from a single aircraft pass is being

* The authors would like to acknowledge the work of Dr. Jack C. Rogers in developing the curvilinear regression technique described in this paper.

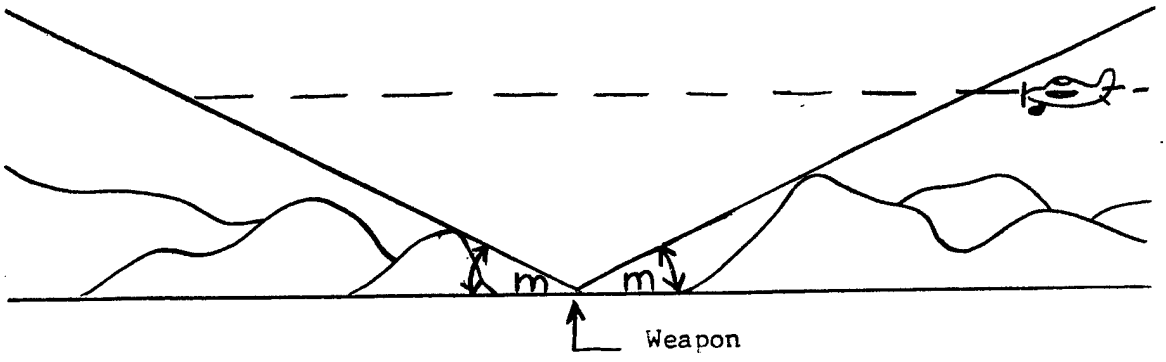


Fig. 1

Terrain Mask Angle

analyzed in detail by hand, while at the same time computer programs are being written which will duplicate the hand method automatically. When the computer programs have been tested, the data reduction will start. The reduced data will then be the basis for analysis of the kind described in this paper.

The aircraft vulnerability of interest here might better be called engagement-vulnerability to distinguish it from the target-vulnerability of the aircraft itself. The engagement-vulnerability is defined by a number of quantities shown in the following figure. These are time under fire (T1), response time (T2), rounds fired (simulated) (F), mean miss distance (D), hits (H), and kills (K). Going down the list one probably gets a better definition of engagement vulnerability, but more assumptions must be made in computing them from the experimental data. Of special interest are the cases when T1 is zero or very small. This means that little or no rounds could be fired at the passing aircraft. The response time T2 for a weapon crew to respond to a sudden appearance of an aircraft is of interest in itself and probably not a good measure of vulnerability. At any rate the six quantities shown in Fig. 2 are the results we want for every wepaon and every set of experimental conditions.

AIRCRAFT ENGAGEMENT-VULNERABILITY	
Per Weapon - Per Ground-Air Engagement	
Dependent Variables	
T1	Time under fire
T2	Response time of crews
F	Rounds fired
D	Mean miss distance
H	Hits
K	Kills

Fig. 2

The quantities in Fig. 3 were varied systematically during the course of the experiment. The number of levels for each variable is also shown.

OPERATIONAL VARIABLES Aircraft and Ground Troops Operating Conditions	
Independent Variables	
V	Velocity of aircraft (4)
A	Altitude (3)
M	Mask angle of terrain (4)
R	Crossing range (2)
W	Warning (2)

Fig. 3

There were a number of factors in the nature of parameters of fixed conditions, not subject to systematic variation and study during the experiment. These were weapon types and troops, along with aircraft evasive action, identification, and target-vulnerability. There were a fixed set of weapon types, while the ground troops were all given the same training, and were rotated around the various ground positions. The aircraft always flew a straight and level path (from different directions, however) and the troops were not required to identify the aircraft before firing.

SOME TECHNIQUES. For exploratory purposes, the effect of aircraft velocity on vulnerability is graphed as shown in Fig. 4.

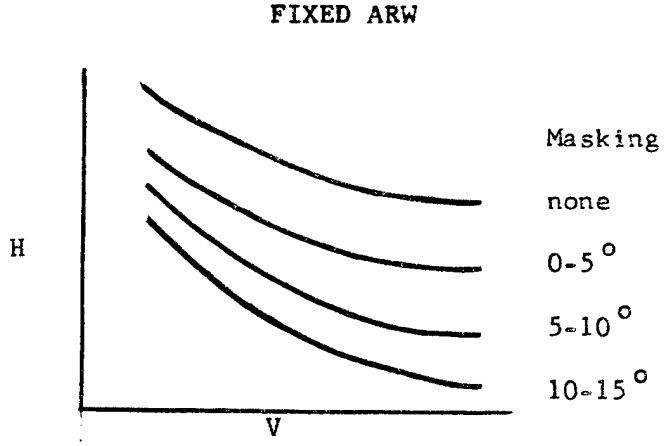


Fig. 4
Velocity Graph

The effect of altitude is graphed as shown in Fig. 5

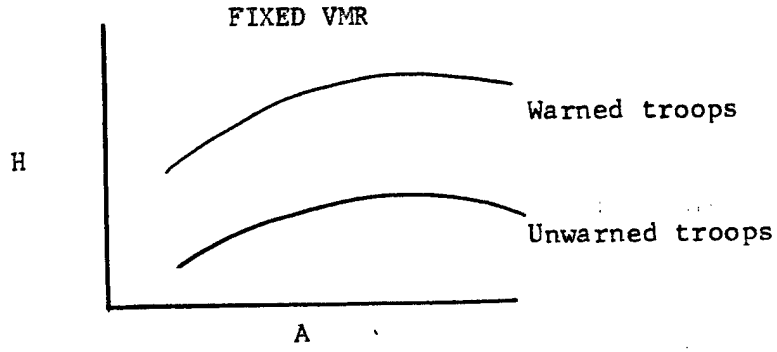


Fig. 5
Altitude Graph

The presence of large fluctuations might already be revealed in the position of the data points on these curves.

Using a regression analysis, it is planned to find all the coefficients F_{ijk} in a polynomial which gives each of the dependent variables (result variables) such as kills (K) or kill probability, as a function of the multilevel variables velocity (V), altitude (A), and mask angle (M). The two-level variables crossing range (R), and alert status (W) will be parameters for the regression. The number of terms in the polynomial increases rapidly with the degree of fit (n), going from 8 to 27 to 64, as n goes from 1 to 2 to 3. The basic method is that of polynomial approximation by orthogonal polynomials and equally spaced points, described by Milne. This requires only a slight adjustment of the experimental data since V, A, and M were designed to have equally spaced points, but were not quite equally spaced due to such difficulties as keeping a light aircraft on a level course. The basic method is repeated a number of times, once for each independent variable until all the coefficients are determined. This process is illustrated in Fig. 6, for n=1. The dependent variable K is expressed as a linear function of V for all possible values of A and M. The resulting (intermediate) coefficients are hence functions of A and M. These coefficients are next expressed as linear functions of A, with the resulting coefficients being functions of M. These latter coefficients are then expressed as linear functions of M. Now by working backwards and substituting the explicit linear form for each coefficient into the previous expressions, the complete polynomial for K is obtained.

CURVILINEAR REGRESSION

$$K = \sum_{ijk=0}^{n} F_{ijk} V^i A^j M^k$$

n	terms
1	8
2	27
3	64

Fixed RW

$$K = F_0(A,M) + F_1(A,M)\, V$$

$$F_0(A,M) = F_{00}(M) + F_{01}(M)\, A$$

$$F_1(A,M) = F_{10}(M) + F_{11}(M)\, A$$

$$F_{00}(M) = F_{000} + F_{001}\, M$$

$$F_{01}(M) = F_{010} + F_{011}\, M$$

$$F_{10}(M) = F_{100} + F_{101}\, M$$

$$F_{11}(M) = F_{110} + F_{111} M$$

Optimum regions

Sensitivity analysis

Error analysis

Fig. 6

In regard to the goodness of fit, a sum of squared deviations is obtained and this will allow an estimate of the amount of randomness involved. The polynomial resulting from this method is adaptable to finding maximum and minimum regions for aircraft vulnerability, a very useful result.

Figure 7 shows a table in which the dependence of vulnerability (in this case hits) upon aircraft crossing range is tested. There are N rounds being considered, and these are classified according to the hits H and non-hits $\overline{H}$, and also according to the rounds fired R at an aircraft flying a path at 300 meters crossing range and the rounds $\overline{R}$ fired at an aircraft flying a path which leads directly over the weapon position. Crossing range has no influence on hits provided it is found that the proportion of hits and non-hits are the same for paths at a crossing range as it is for all rounds generally. Fluctuations of this proportion due to pure chance will also be considered.

SIGNIFICANCE TESTS

(Contingency Tables)

	R	$\bar{R}$	
H	(HR)	$(H\bar{R})$ →	(H)
$\bar{H}$	$(\bar{H}R)$	$(\bar{H}\bar{R})$ →	$(\bar{H})$
	↓ (R)	↓ $(\bar{R})$ →	N

Aggregated V-A-M-W

H Independent of R

Provided

$$\frac{(HR)}{(\bar{H}R)} = \frac{(H)}{(\bar{H})}$$

Fig. 7

Figure 8 shows a correlation table containing plus signs for positive correlation coefficients, minus signs for negative ones, and question marks for cases where even the sign of the coefficient is in doubt at our present stage of knowledge.

LINEAR CORRELATIONS

Independent Variables

DEPENDENT VARIABLES	V	A	M	R	W
T	-	+	-	+	+
F	-	+	-	+	+
D	+	-	+	-	-
H	-	?	-	-	+
K	-	?	-	-	+

$$C = \frac{\sum XY}{\sigma_x \sigma_y}$$

Fig. 8

In the application of the analysis of variance to the experimental data, the lack of a second replication requires special consideration. It was generally not the case that the experimental conditions (treatments) were repeated, although a few were run more than once. This means that there is only one result (yield) for every condition, or one vulnerability measure for each combination of the independent variables V, A, M, R, and W. With these variables there are 4 x 3 x 4 x 2 x 2, or 192 different conditions. In order to introduce some variation for analysis, a number of approaches suggest themselves. First, if it is known from preliminary analysis that some variable such as crossing range (R) has a negligible effect upon vulnerability, this effect will not be investigated and consequently will be randomized. This gives 4 x 3 x 4 x 2 or 96 different conditions, each condition now having two results. A complete factorial with two replications is thus obtained. Other approaches are based on a lack of sensitivity of vulnerability to a multilevel variable such as altitude (A), or mask angle (M). In this case either A or M could be reduced to two levels, called high and low altitude, or high and low mask angle. In either case, a complete factorial with two replications would be obtained. With these two replications a completely randomized design is applicable with 95 degrees of freedom for the variation due to error, 1 for replications, and 95 for main effects and interactions.

Another approach is to ignore the highest order interaction, namely V with A, M, R, and W, and treat it as an error term. In this case, there is one replication with 173 degrees of freedom for the main effects and interactions, and 18 for experimental error.

The results of these various approaches is to permit a test of the hypothesis that the mean result for each condition is the same. Since we are rather confident that they are not the same, it will then be desirable to find out which conditions might be equivalent and which of then have the largest effect on vulnerability.

SUMMARY. Some techniques planned for the analysis of results from the ORO part of the Aircraft Vulnerability Experiment have been described, along with some background information on the experiment and a statement of the data analysis accomplished to date.

DESIGN FOR A PROPOSED FIELD EXPERIMENT WITH LIGHT AT WEAPONS

R. E. Tiller, J. D. Reed, J. P. Young
Operations Research Office, the Johns Hopkins University

OBJECTIVES. The experiment described here was designed to determine accuracies under simulated tactical conditions of a family of current and prototype shoulder-fired AT* weapons, and to develop data which will serve as guide-lines for research and development leading toward optimized weapons.

BACKGROUND. AT warfare places stringent requirements on hand-held infantry weapons. It is essential that these be light in weight, rugged, simple to operate and maintain with minimum training, and most important, they must offer a high probability of hitting and killing the tank with the first round. Tanks very seldom work singly, and even when in sections or platoons usually are accompanied by protecting infantry. Disclosure by fire is therefore a serious problem to the AT gunner; he can fire only one round, and must then move quickly to another position, or be killed. This problem emphasizes the need for a first-round hit.

Hit probability is influenced by a number of factors which may be arbitrarily divided into "ballistic" and "gunner" errors. Even on the training range at known distances, firing at clearly visible targets, the gunner's error greatly exceeds the error of the weapon and ammunition, and in the operational situation, we can in nearly every instance ignore the ballistic factor. If a weapon is acceptable to the Army, the problems associated with ballistic engineering are in most cases minor compared to those introduced by the gunner. Velocity will be acceptably constant, the shot group at a given range will meet the stated specifications, and the weapon will be satisfactorily rugged.

The gunner's errors are another thing, however. His errors in aiming, in canting the weapon, and most important to the shoulder-fired AT weapons, (which have characteristically low velocities), in range estimation, are of primary importance in the system accuracy.

Interaction of velocity with range estimation error therefore determines to a large extent, the probability of the infantryman hitting the enemy tank.

TEST ITEMS. The weapons to be tested are the cuurent 3.5 Rocket Launcher (M20); a new prototype rocket launcher which will be fired in two ways, a new recoilless rifle, this last weapon to be tested with 3 fire control systems and the cal. 30 rifle (M1) which will simulate a hypothetical flat trajectory high velocity weapon.

Theoretical analysis of hit probability as a function of range indicates an elliptical normal distribution, due to the large vertical component of error introduced by range estimation. To check the equation which will be used for determining P_H and determine the constants in that equation, it will be necessary to test the weapons at a number of ranges appropriate to each weapon. A large number of firings will be required, determining the error associated with each firing.

---

* Antitank

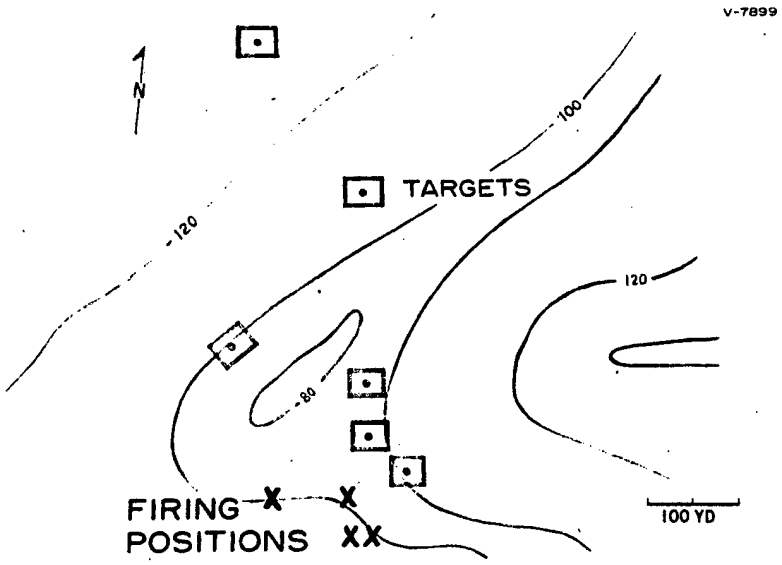


Fig. 1. Hypothetical layout of firing positions and targets for Phase I.

Test Subjects and Training. Fifty-two soldiers, preferably men with no particular M.O.S., who have recently completed Basic Infantry Training, and selected to eliminate critical physical differences, will serve as test subjects. Twenty-six will be trained as gunners and 26 as loaders. Prior to the experiment, after receiving instruction on all test weapons, they will be assigned to four groups of 6 gunner-loader teams (plus spares).

PROCEDURE. It will be necessary to conduct three separate experimental phases to achieve all of the desired objectives.

The first will be a determination of hit probability against static targets, for these we will use 7½ ft square O.D. panels, electro-mechanically controlled to permit exposure in the desired sequence.

The second will employ moving targets ("buttoned-up" tanks on a prescribed course using accepted evasive tactics). The last phase will deal with the determination of second round hit probability as influenced by the first round, and will again employ the panel targets used in Phase I.

TARGET AREAS. Two similar target areas will be required for Phase I of the experiment. A hypothetical target layout is shown in Figure 1.

On these areas the 7½ ft panel targets will be placed at ranges where .25, .50, and .75 hit probabilities are expected on the basis of theoretical calculations; some ranges will overlap and permit the use of the same targets for different weapons.

To simulate more closely actual operating conditions, a series of explosive charges will be detonated near the firing points and near the targets.

EXPERIMENTAL DESIGN. The test schedule is designed to minimize specific learning of ranges. The subject will fire the M-1 rifle and one other weapon in an area each half day. The locations of the firing points and targets will be changed, so that the subjects will not be able to transfer specific information from one situation to another. The order of appearance of the targets will also be varied. As a result, the subjects will gain only a general knowledge of the target area, similar to what might be expected in an operational situation. Furthermore the subjects will not be able to profit by talking to other men who have just completed firing, because each order faces a slightly different situation, and no firers will be permitted to enter the target area at any time.

PHASE I--STATIC TARGETS. The experimental factors to be investigated are:

24 men
2 firing positions
7 weapons or weapon combinations
3 ranges for AT weapons; all ranges for the M1 rifle

Each will be systematically varied in a balanced experimental design, utilizing the 24 gunners (with loaders as required) in four groups of six

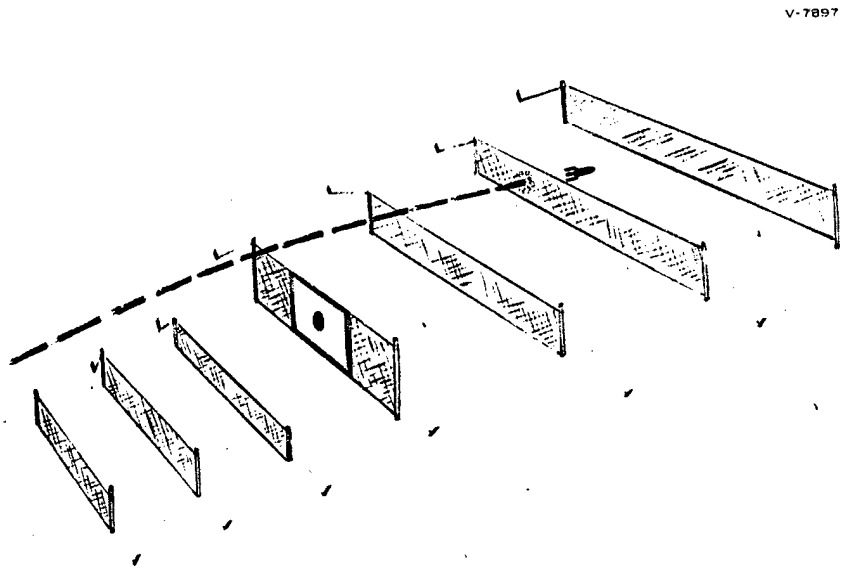


Fig. 2. Diagram of target with associated screens designed to determine "miss distances."

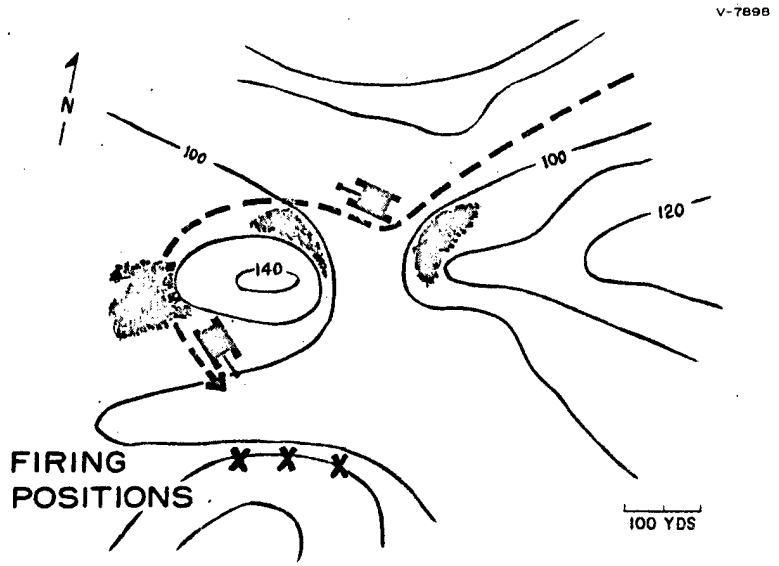


Fig. 3. Hypothetical layout of firing positions and terrain features required for Phase III.

who will fire individually, using one type of weapon in each area in each time period, with the exception of the M-1 which will be fired at all targets.

Weapons fired and their ranges (close, mid-range or distant) will be systematically varied among the firers as well as among firing position, area and time of day.

Twenty-four men, firing in two positions, will yield 48 pieces of information for each range and inasmuch as aiming error expresses itself in angular error, this procedure may offer data for one range which supports the other two ranges.

At a minimum, this design will yield 48 pieces of data at a given range which even if unrelated to the other ranges, will provide satisfactory accuracy for our determination of hit probability.

Adequately complete data for our determinations of P_H will involve not only the distribution of hits on the 7½ ft panels, but will require also measurement of the dispersion of at least 95% of the rounds which fall short or pass over the target. A fully satisfactory and effective technique has not yet been determined, but this problem is currently under study.

The feasibility of installing a series of vertical wire screens appropriately located to the front and rear of the target position is one of the suggested methods. Figure 2 is a diagrammatic representation of a typical target with its associated screens. The height, width and location of the screens is calculated to accomodate the weapon having the sharpest angle of fall at the given range.

Preliminary field testing indicates that although this system does not pose any problem by obscuring the target or by detracting seriously from the tactical realism of the target area (when painted O.D., the fences are invisible at ranges greater than 75 yards), they are highly vulnerable to damage by ricochets.

A test is scheduled in the near future to evaluate the accuracy with which misses can be located by observing the point of impact through a B.C. scope, and determining by standard surveying techniques the horizontal and vertical displacement of the round in relation to the target.

PHASE II. In this exercise a tank will cover a prescribed course, unknown to the gunner, and will use accepted evasive tactics (changes of speed and direction). Three weapons will be used, each at two ranges.

As indicated in Figure 3, terrain will provide a feature allowing the tank to appear for 20 seconds at a 50% P_H range for two of the test weapons. It will then disappear, and reappear at 50% P_H range for the third, shorter range weapon. Each weapon will fire one round at each range.

From these data on weapons of comparatively high and low velocities, we hope to evaluate effectiveness of intermediate velocities.

No difficulty for the observers is anticipated in identifying the performance of the weapons, since each has a characteristic velocity and trajectory.

As in Phase I, two areas will be needed. Two groups will fire Phase II while the others are firing Phase I.

PHASE III--SECOND-ROUND HIT PROBABILITY. Determination of improvement in P_H on firing a second shot will require the use of live ammunition for the first round for realistic sensing.

The areas described for Phase I will be used. Two weapon combinations will be employed.

Each man will fire two rounds at each of five targets. The two weapons are so dissimilar that they may be fired concurrently. The target sequence will be varied, but only one firing position will be used. As in Phase I, each group will fire in one area for one period to achieve balance.

It is our hope that this proposed field test will not only yield realistic values for P_H with existing shoulder-fired AT weapons and for several prototype weapons, but will indicate with validity not obtainable from engineering tests, the critical parameters for designing the best weapons in this category.

UNBIASED ESTIMATION BASED ON TRANSFORMED VARIABLES, WITH PARTICULAR REFERENCE TO CLOUD SEEDING EXPERIMENTS*

Jerzy Neyman and Elizabeth L. Scott
University of California, Berkeley

1. INTRODUCTION. Because of the notorious skewness of the distribution of meteorological observations, their statistical analysis frequently begins with a transformation of variables. Thus, for example, instead of dealing with amounts of precipitation measured in inches or with runoff measured in acre-feet, one works with the square roots of these quantities, or perhaps with their logarithms, etc. With a certain amount of good luck, the distribution of the transformed variables approaches the normal distribution and also satisfies certain other conditions. As a result, a reliable analysis may be performed using the standard statistical techniques, which were developed on certain restrictive assumptions.

Granting that the transformation chosen is satisfactory, certain parts of the analysis can be performed entirely in terms of the transformed data, without any reference to the natural units (inches, etc.) in which the original observations are expressed. However, this is not true with problems of point estimation of the parameters. For example, in order to be intelligible, the estimates of the average increase in precipitation or in runoff ascribable to seeding must be expressed in units appropriate to these quantities, not in square roots or in logarithms. The customary procedure for obtaining estimates in the original units is to calculate estimates in the transformed units and then transform these backwards (by squaring or by taking antilogarithms, etc.). However, it happens that the backward transformation applied to an unbiased estimate leads, generally, to a biased estimate. Occasionally, this bias is trivial and is overshadowed by the uncertainty of the estimate due to random fluctuations. In other cases, the bias is quite important. Furthermore, even if the bias of an estimate is small compared to its random error, the bias may acquire importance when this estimate is combined with other biased estimates derived from other sets of data.

A case in point is the combined evaluation of a number of cloud seeding experiments. Because of the great variability of the observations and the relative scarcity of data, the estimated increase in precipitation obtained in any particular experiment ordinarily is shaky. Thus, it appears reasonable to try to combine the results of a number of experiments so as to obtain an estimate of the average effect of seeding. If the estimates calculated for particular experiments are all unbiased, then the combined unbiased estimate is easy to obtain by calculating the appropriate weighted average. On the other hand, the averaging of biased estimates, particularly if the bias always has the same sign, may lead to serious errors: the random errors attached to the single esitmates will tend to average out, but not the bias. In fact, by examining a recently published combined evaluation of a number of commercial cloud seeding operations, the authors found that, due to the bias involved in the

* This paper was prepared with the partial support of the Office of Ordnance Research, U. S. Army under Contract DA-04-200-ORD-171, Task Order 3.

estimates, the average effect ascribed to seeding is likely to have been overestimated by a factor of two.

The purpose of the present paper is to provide formulas for the unbiased estimation of parameters in their natural units when the analysis is based on transformed variables. While we are primarily concerned with the evaluation of cloud seeding experiments, and, more specifically, with unbiased estimation of the effects of seeding, the same formulas are likely to be useful in other cases.

2. NOTATION AND ASSUMPTIONS. In order to estimate the increase in precipitation ascribable to seeding, it is necessary to estimate the precipication which would have fallen in the target in the absence of seeding. A common method uses the precipitation in one or more, say s, comparison areas presumed to be free of any effect of seeding and so considered as control areas. Then regression analysis provides the estimate needed.

The observations may be the amounts of precipitation, measured in inches, falling in the areas considered during specified intervals of time, or the amounts of runoff. The intervals of time may be variously defined "storms," twelve-hour periods, days, months or years. The exact nature of the observations and the particular intervals of time to which they refer are irrelevant to the discussions which follow. For this reason, and for the sake of simplicity in wording, we shall speak of the amounts of precipitation from a storm.

The observations will consist of a certain number m of seeded and a certain number n of not-seeded storms, and we shall assume that both groups represent random samples from the same well-defined population of storms. We need symbols to denote the amounts of precipitation in the target and in the controls, first, generally, for a storm of the particular category and then for the j-th storm of the available sample. For a seeded storm in general, the amounts of precipitation in the s controls will be denoted by

$$(1) \qquad {}^*X_1, {}^*X_2, \ldots, {}^*X_s,$$

respectively. Occasionally, it will be convenient to use one symbol to denote these s variables; we shall use the symbol ${}^*X_\cdot$ for this purpose, so that ${}^*X_\cdot = ({}^*X_1, {}^*X_2, \ldots, {}^*X_s)$. The corresponding seeded precipitation in the target will be denoted by ${}^*Y$. All this applies to a seeded storm "in general." When referring to the j-th seeded storm of the available sample, the corresponding symbols will be

$$(2) \qquad {}^*X_{\cdot j} = ({}^*X_{1j}, {}^*X_{2j}, \ldots, {}^*X_{sj}) \quad \text{and} \quad {}^*Y_j,$$

respectively. Here, then, $j = 1,2,\ldots,m$. The notation for the not-seeded storms, either generally or for the j-th member of the available sample, will be the same except that we shall omit the asterisks. Thus, for example, Y_j will denote the target precipitation from the j-th unseeded storm.

Symbols involving the letters X and Y will denote the amounts of precipitation measured in the original units, that is, in inches, etc. We now introduce a corresponding set of symbols to denote the transformed variables, replacing X by U and replacing Y by V. In this way, for example, $*V_j$ will mean the target precipitation in the transformed units from the j-th seeded storm.

In this paper we are not concrrned with the choice of the function for transforming from the original variables X or Y to the transformed variables U or V. The literature on this subject is extensive. We are interested in the transform back into the original units. For this reason, it is convenient to denote the function carrying the original variables into the transformed variables by f^{-1}, perhaps with subscript, while the inverse transformation carrying the transformed variables back into the original variables will be denoted by f, with appropriate subscript. It will be seen that since our interest is reversed, the notation is also reversed. For short, we call f the transforming function.

Thus, it will be assumed that each of the original variables X or Y is functionally related to the corresponding transform, and that this relation is the same for seeded and for not-seeded storms. In other words, we postulate the existence of s+1 monotone functions f_0, f_1, f_2, ..., f_s such that

$$(3) \qquad *Y = f_0(*V) \quad \text{and} \quad Y = f_0(V)$$

and, for $i = 1,2,\ldots,s$

$$(4) \qquad *X_i = f_i(*U_i) \quad \text{and} \quad X_i = f_i(U_i).$$

Frequently, there s+1 functions all coincide, in which case the identifying subscripts are superfluous.

As stated in the Introduction, the subject of this paper is limited to the estimation of certain parameters. For this reason it will be assumed throughout that the many pitfalls involved in the evaluation of cloud seeding experiments are successfully avoided and, in particular, that the transformed variables U. and V corresponding to not-seeded storms satisfy exactly the following condition (i) and that the variables *U. and *V corresponding to seeded storms satisfy either condition (ii) or condition (iii):

(i) For any not-seeded storm with transformed precipitation U. in the control areas equal to $u_\cdot = (u_1,u_2,\ldots,u_s)$, the transformed target precipitation V is normally distributed with mean

$$(5) \qquad E(V|U_\cdot=u_\cdot) = \alpha_0 + \sum_{i=1}^{s} \alpha_i u_i = \mu(u_\cdot), \quad \text{say,}$$

and with a fixed variance σ^2, independent of $u_{\cdot}$. The variance σ^2 will be called the residual variance. Here, the α and σ are unknown constants.

We shall use two alternative conditions, say (ii) and (iii), regarding the transformed precipitation from seeded storms. Then each combination, (i) with (ii) on the one hand and (i) with (iii) on the other, will serve to specify a separate problem of estimating the effects of seeding.

(ii) Nothing is assumed regarding $*U_{\cdot}$ and $*V$ except that, for each value $*u_{\cdot}$ of $*U_{\cdot}$, the variable $*V$ has a finite mean $E(*V|*u_{\cdot})$, which may or may not equal $E(V|*u_{\cdot})$ in (5).

(iii) For each possible set of precipitation amounts $*u_{\cdot}$ in the control, the transformed precipitation $*V$ in the target is normally distributed about a mean,

$$(6) \qquad E(*V|*u_{\cdot}) = \beta_0 + \sum_{i=1}^{s} \beta_i *u_i = *\mu(*u_{\cdot}), \quad \text{say,}$$

with a fixed variance $*\sigma^2$. Here the β represent constant coefficients which may or may not equal the α in (5). Also, the residual variance $*\sigma^2$ may but need not equal the residual variance σ^2.

3. SPECIFICATION OF THE PROBLEM. As we have said, the problem of unbiased estimation of the effects of seeding has somewhat different specifications according to whether we assume condition (ii) or the more restrictive condition (iii).

The hypothesis basic for the evaluation is, of course, condition (i). For any preassigned conditions in which the transformed control precipitation has values $u_{\cdot} = (u_1, u_2, \ldots, u_s)$ this assumption determines the distribution of the corresponding transformed target precipitation V to be observed without seeding. The expectation of this not-seeded precipitation in the target expressed in the original units is simply, say,

$$(7) \qquad \theta(u_{\cdot}) = E\{f_0(V)|u_{\cdot}\} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} f_0(V) e^{-\frac{1}{2\sigma^2}\left[V-\mu(u_{\cdot})\right]^2} dV.$$

If the coefficients α and the variance σ^2 were known, the expectation (7) could be computed easily. However, these constants are unknown and the best that can be done is to use the n not-seeded storms to obtain unbiased estimates of them. The formulas leading to these estimates are given in a number of textbooks, for example, in [1]. For the two most important cases s=1 and s=2, these formulas are reproduced in Section 6. Using the estimates $\hat{\alpha}_i$ of α_i we obtain the unbiased estimate

$$(8) \qquad \hat{\mu}(u_{\cdot}) = \hat{\alpha}_0 + \sum_{i=1}^{s} \hat{\alpha}_i u_i$$

of (5). The estimate $\hat{\mu}(u_\cdot)$ is known to be normally distributed about $u(u_\cdot)$ and has variance $\sigma^2(u_\cdot) = \lambda^2(u_\cdot)\sigma^2$, where $\lambda^2(u_\cdot)$ is a known function of $u_\cdot$. In general, $\lambda^2(u_\cdot)$ is a quadratic in $u_\cdot$, which attains its minimum value of $1/n$ when $u_\cdot$ is equal to the average amounts of transformed control precipitation from not-seeded storms and then increases without limit when $u_\cdot$ diverges from these averages.

Also, the same familiar normal theory implies that the sum of squares of residuals

$$(9) \qquad S^2 = \sum_{j=1}^{n} (V_j - \hat{\alpha}_0 - \sum_{i=1}^{s} \hat{\alpha}_i u_{ij})^2$$

is a statistic independent of $\hat{\mu}(u_\cdot)$ and, when divided by σ^2, is distributed as χ^2 with $\nu = n-s-1$ degrees of freedom. As a result, the quotient S^2/ν is an unbiased estimate of the residual variance.

As we shall see below, the basic problem of evaluating a cloud seeding experiment consists in using the two statistics $\hat{\mu}(u_\cdot)$ and S^2, computable from data on not-seeded storms, in order to obtain an unbiased estimate of $\theta(u_\cdot)$ as defined by (7).

We turn now to the conditions (ii) and (iii) regarding the seeded storms. Condition (ii) does not imply any link between the seeded target precipitation from any two storms. Thus, for example, (ii) admits the possibility that the effect of seeding one type of storm may be positive and that of seeding another negative. Further, under condition (ii) it is possible to draw conclusions as to the seeded target precipitation only in the situations prevailing during the storms that actually were seeded. For this reason, under condition (ii), the evaluation of the experiment must be reduced to comparing the average actual seeded target precipitation and the average expected not-seeded target precipitation in the conditions of the m actual seeded storms. In other words, the quantity to be estimated is

$$(10) \qquad \frac{1}{m}\sum_{j=1}^{m} E(*Y_j) - \frac{1}{m}\sum_{j=1}^{m} \theta(*u_j).$$

The only unbiased estimate of the first term in (10) is the simple average

$$(11) \qquad \frac{1}{m}\sum_{j=1}^{m} *Y_j = *\bar{Y} \quad .$$

In order to estimate the second term in (10), we need a general formula for estimating (7).

Under condition (iii) the situation is somewhat more flexible. Here the evaluation of the experiment need not be restricted to the amounts

observed in the storms actually seeded, which may happen to be atypical. For example, the evaluator may fix in advance arbitrary amounts of control precipitation $u_\cdot$, perhaps representing "normal" amounts per storm observed over a number of years, and then estimate from the data the expected effect of seeding. Now the quantity to be estimated is the difference

$$(12) \qquad {}^*\theta(u_\cdot) - \theta(u_\cdot),$$

where ${}^*\theta(u_\cdot)$ is defined like $\theta(u_\cdot)$ in formula (7), namely,

$$(7a) \qquad {}^*\theta(u_\cdot) = E\{f_0({}^*V) \mid u_\cdot\} = \frac{1}{{}^*\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} f_0(V) e^{-\frac{1}{2{}^*\sigma^2}\left[V - {}^*\mu(u_\cdot)\right]^2} dV.$$

The problem of estimating ${}^*\theta(u_\cdot)$ is the same as that of estimating $\theta(u_\cdot)$ except that the data on seeded storms are used.

In conclusion, then, whether we adopt hypothesis (ii) or (iii), the problem of evaluating the experiment requires the formula for an unbiased estimate of a quantity of type (7).

4. TRADITIONAL SOLUTION AND ITS BIAS. The traditional method of estimating $\theta(u_\cdot)$ consists in computing $\hat{\mu}(u_\cdot)$, the estimated expected transformed target precipitation without seeding, in conditions such that the transformed control precipitation is $u_\cdot$, and then applying the backward transformation. In other words, the traditional estimate of $\theta(u_\cdot)$ is $f_0[\hat{\mu}(u_\cdot)]$. We now show that, with the usual transforming functions, this estimate is biased. This means that its expectation is not equal to the quantity $\theta(u_\cdot)$ to be estimated. There is a systematic error so that even with an infinite number of observations the estimate would not equal the true value. Also, if a number of such estimates are averaged, the bias will not tend to average out.

It is well known that the conditional distribution of the precipitation Y in the target, given a specified amount of precipitation X_i in the i-th control, will generally have a variance that increases with an increase in X_i. One of the main purposes of the transformation of variables is to stabilize this conditional variance. There are many functions f^{-1} which will accomplish this; some are used in one case, some in another. But, in order to be useful, the functions f^{-1} must "shrink" the larger observations. Hence, the transforming function f must be concave. In other words, as illustrated in Figure 1, the graph of the transforming function f is a curve that lies entirely above a tangent straight line, no matter where on the curve this tangent is drawn.

Figure 1 shows three of the transformations used in the evaluation of cloud seeding experiments. The first panel corresponds to the square root transformation, that is, to the case $Y = f_0(V) = V^2$. The second panel corresponds to the logarithmic transformation so that

$Y = f_0(V) = 10^V = e^{kV}$, where k denotes the natural logarithm of 10. Finally, the third panel corresponds to the equiprobability gamma transformation advocated by Thom [2]. Here $Y = f_0(V)$ is defined by the relation

$$(13) \qquad \frac{1}{\beta^\alpha \Gamma(\alpha)} \int_0^Y t^{\alpha-1} e^{-t/\beta} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^V e^{-t^2/2} dt,$$

where α and β are certain positive constants.

Now we show that the traditional estimate is biased. We begin by examining the relation between the quantities $\theta(u.)$ and $f_0[\mu(u.)]$. As formula (7) indicates, $\theta(u.)$ is the expectation of the variable $Y = f_0(V)$. Thus, $\theta(u.)$ is the weighted average of all possible values of $f_0(V)$, each value weighted by the probability of that value. For each value of V the corresponding value of $Y = f_0(V)$ is equal to the ordinate of a curve similar to those in Figure 1. The difficulty arises because V is not perfectly determined by the observed precipitation u. in the controls; V is a random variable normally distributed with mean $\mu(u.)$.

Let us compare the tangent at the ordinate $f_0[\mu(u.)]$, corresponding to the abscissa $\mu(u.)$, with the curve itself. The equation of this tangent is, say,

$$(14) \qquad Z(V) = f_0[\mu(u.)] + [V-\mu(u.)]f_0'[\mu(u.)] .$$

The weighted mean of Z(V), subject to the variability of V, is $f_0[\mu(u.)]$ since the mean of $V-\mu(u.)$ is zero. However, since the curve is above the tangent, the weighted mean, with the same weights, of its ordinates $f_0(V)$, namely $\theta(u.)$, must be greater than the mean $f_0[\mu(u.)]$ of the tangent.

Incidentally, the occrurence of the difference between the two means does not depend on the normality of V but persists irrespective of the weights used in averaging. For example, the simple arithmetic mean of the two numbers $x_1 = 2$ and $x_2 = 4$ is $\bar{x} = 3$. The square of this average is $\bar{x}^2 = 9$ while the average of the squares of the same numbers is $(2^2+4^2)/2 = 10$ which is larger than $\bar{x}^2$. There is another conclusion which can be drawn heuristically from the above discussion: The greater the variability of V, that is, the greater its variance σ^2, the greater the difference between $\theta(u.) = E[f_0(V)]$ and $f_0[\mu(u.)]$. This results from the simple remark that, with an increase in the variance σ^2, values of V substantially different from the mean $\mu(u.)$ and, therefore, values for which the difference between $f_0(V)$ and Z(V) is large will have greater weight. This is true in general. For example, the two numbers $x_3 = 1$ and $x_4 = 5$ have the same average $\bar{x} = 3$ as the numbers x_1 and x_2 above. However, x_3 and x_4 have a greater spread. As a result, the mean of their squares $(x_3^2 + x_4^2)/2 = 13$ is greater than the mean 10 of the squares of x_1 and x_2.

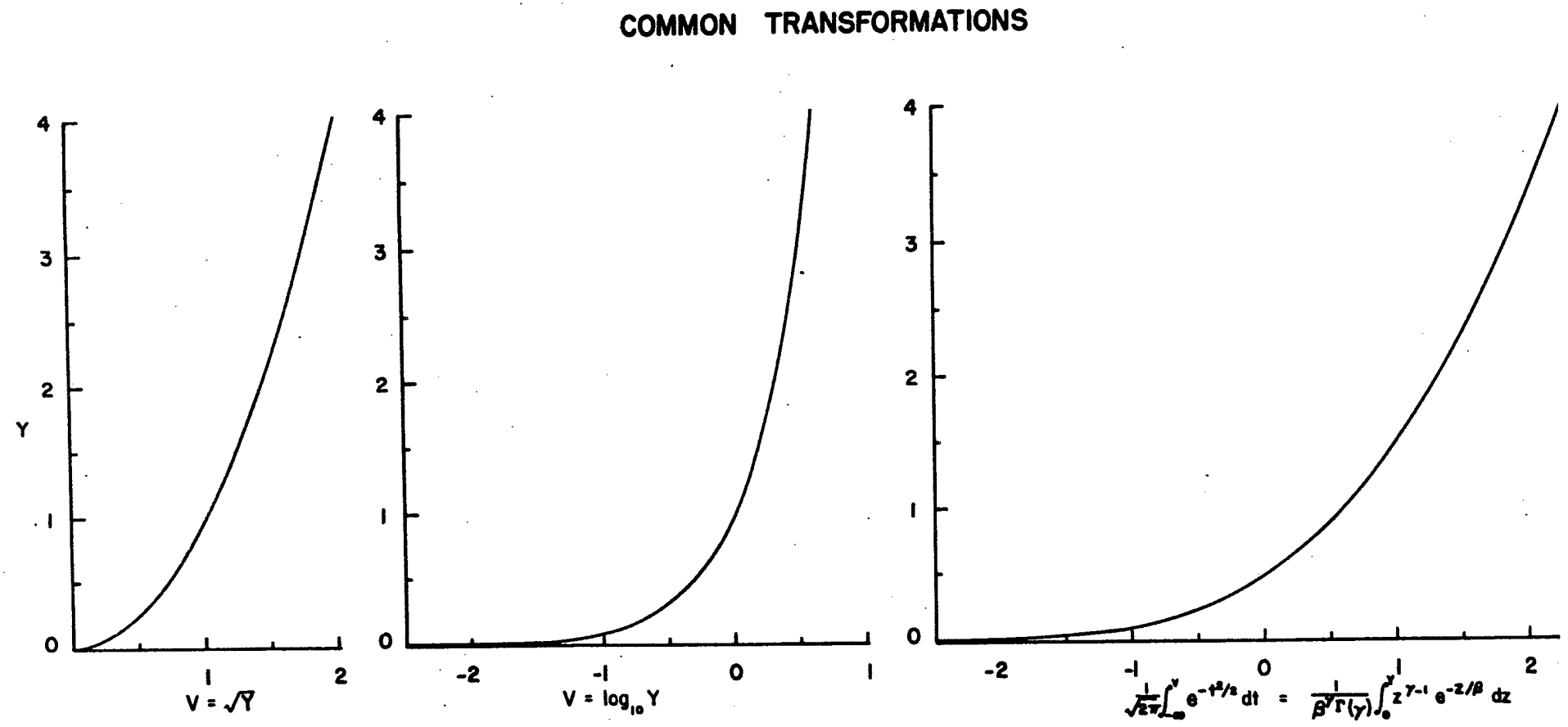

COMMON TRANSFORMATIONS
Y
V = √Y
V = log10 Y
$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\gamma} e^{-t^2/2}\,dt = \frac{1}{\beta^{\gamma}\Gamma(\gamma)}\int_{0}^{\gamma} z^{\gamma-1} e^{-z/\beta}\,dz$

Figure 1

The two conclusions, (a) that the mean of a concave function is greater than the value of this function taken at the mean of its argument, and (b) that the difference between the mean of a concave function and the value of the junction at the mean increases with an increase in the variability of the argument, jointly explain the reason for the bias in the traditional estimate $f_0[\hat{\mu}(u.)]$.

As we have said, $\hat{\mu}(u.)$ is a normal variable with mean $\mu(u.)$ and variance $\lambda^2(u.)\sigma^2$. The mean of V is also $\mu(u.)$ while its variance is σ^2. Ordinarily u. will not differ very much from the averages of the transformed control precipitation in not-seeded storms. In this case, the value of $\lambda^2(u.)$ is less than unity and so the variance of $\mu(u.)$ will be less than that of V. Consequently,

$$(15) \qquad f_0[\mu(u.)] < E\{f_0[\hat{\mu}(u.)]\} < E\{f_0(V)|u.\} = \theta(u.),$$

so that $f_0[\mu(u.)]$ will systematically underestimate $\theta(u.)$. On the contrary, if u. is far away from the average transformed amounts of control precipitation of not-seeded storms, then $\lambda^2(u.) > 1$, and we have

$$(16) \qquad f_0[\mu(u.)] < E\{f_0(V)|u.\} = \theta(u.) < E\{f_0[\hat{\mu}(u.)]\},$$

so that the traditional estimate $f_0[\mu(u.)]$ will tend to overestimate $\theta(u.)$ in this case.

5. GENERAL FORMULAS FOR THE UNBIASED ESTIMATE OF $\theta(u.)$. In a recent paper [3] we considered in some detail the problem of a minimum variance unbiased estimate of a quantity of the type of $\theta(u.)$. In the present section we give without proof two general formulas applicable to a broad class of transforming functions f_0. Before proceeding to details we interpose two remarks.

Remark 1. As we saw in Section 4, the expected value of V, namely $\mu(u.)$, does not by itself determine the value of $\theta(u.)$; the variance σ^2 plays an important role. As a result, it is clear that an unbiased estimate of $\theta(u.)$ cannot depend solely on the estimate $\hat{\mu}(u.)$ of $\mu(u.)$ but must involve the statistic S^2 which serves to estimate σ^2. Consequently, any unbiased estimate of $\theta(u.)$ will be a function, say $\hat{\theta}[\hat{\mu}(u.), S^2]$ of the two arguments $\hat{\mu}(u.)$ and S^2.

Remark 2. The two statistics $\hat{\mu}(u.)$ and S^2 form a so-called "sufficient" system for the parameters $\hat{\mu}(u.)$ and σ^2. Also, this sufficient system is "boundedly complete." According to a theorem of Lehmann and Scheffé [4], if a function of these two statistics is an unbiased estimate of $\theta(u.)$ then it is necessaaily the minimum variance unbiased estimate and is the unique estimate of this kind. Consequently, each of the formulae below giving $\hat{\theta}[\hat{\mu}(u.), S^2]$ is the expression for the minimum variance unbiased estimate of the corresponding $\theta(u.)$. Any other formula would either be equivalent to that produced, or give a biased estimate, or give an unbiased estimate with greater variance.

The first formula is of somewhat greater generality than the second. In addition, it is easier to apply for certain types of transforming

functions, although more difficult for other types. In order that the first formula be applicable, it is necessary and sufficient that the transforming function f_0 have derivative of all orders and that the two series

$$(17) \qquad \sum \frac{1}{n!} f^{(2n)}(0) z^n \quad \text{and} \quad \sum \frac{1}{n!} f^{(2n+1)}(0) z^n$$

have infinite radii of convergence. If these conditions are satisfied, the minimum variance unbiased estimate of $\theta(u.)$ is given by

$$(18) \qquad \hat{\theta}\left[\hat{\mu}(u.), s^2\right] = f_0(0) + \sum_{n=1}^{\infty} f_0^{(n)}(0) T_n ,$$

where

$$(19) \qquad T_{2n} = \sum_{k=0}^{n} \frac{(2n)!}{(2k)!(n-k)!} \hat{\mu}^{2k}(u.) \left\{\frac{s^2}{4}\left[1 - \hat{\lambda}^2(u.)\right]\right\}^{n-k} \frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2} + n - k\right)}$$

and

$$(20) \qquad T_{2n+1} = \sum_{k=0}^{n} \frac{(2n+1)!}{(2k+1)!(n-k)!} \hat{\mu}^{2k+1}(u.) \left\{\frac{s^2}{4}\left[1 - \lambda^2(u.)\right]\right\}^{n-k} \frac{\Gamma\left(\frac{\nu}{2}\right)}{\Gamma\left(\frac{\nu}{2} + n - k\right)} .$$

Although formula (18), combined with (19) and (20), looks complicated, it is easy to apply to some of the transforming functions advocated by Tukey [5]. These functions are of the form

$$(21) \qquad f_0(V) = V^p - a,$$

where a is a fixed number and p a positive interger. It will be seen that at $V = 0$ all the derivatives of (21) vanish with the exception of

$$(22) \qquad f_0^{(p)}(0) = p! .$$

Consequently, formula (18) reduces to

$$(23) \qquad \hat{\theta}\left[\hat{\mu}(u.), s^2\right] = -a + T_p ,$$

where T_p has the form (19) or (20) according to whether p is even or odd. The square root transformation is a particular case of (21), with $p = 2$ and $a = 0$. In this case, then, after simplification,

$$(24) \qquad \hat{\theta}\left[\hat{\mu}(u.), s^2\right] = T_2 = \hat{\mu}^2(u.) + \frac{s^2}{\nu}\left[1 - \lambda^2(u.)\right] .$$

The second general formula for the unbiased estimate $\theta\left[\hat{\mu}(u.), s^2\right]$ was deduced for a special category of transforming functions, which we call recursive. The formula is easily applicable to these functions, which are rather common. These functions are characterized by the

differential equation of the second order which they must satisfy, namely,

$$f''(V) = A + Bf(V) \quad , \tag{25}$$

where A and B are two constants. We assume that at least one of these constants must be different from zero; otherwise f would be a linear function in which case it would not stabilize the residual variance.

If the transforming function f is recursive in the above sense, and $B \neq 0$, then the minimum variance unbiased estimate of $\theta(u_.)$ is given by

$$\hat{\theta}[\hat{\mu}(u_.), s^2] = \Phi\{B[1-\lambda^2(u_.)]s^2, \nu\}\{f[\hat{\mu}(u_.)] + \frac{A}{B}\} - \frac{A}{B} \quad , \tag{26}$$

where, generally,

$$\Phi(x,\nu) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu}{2}+n)} \cdot (\frac{x}{4})^n \tag{27}$$

$$= 1 + \frac{x/2}{1!\nu} + \frac{(x/2)^2}{2!\nu(\nu+2)} + \frac{(x/2)^3}{3!\nu(\nu+2)(\nu+4)} + \cdots \quad .$$

When $B = 0$, which is when f is a quadratic function, formula (26) reduces to

$$\hat{\theta}[\hat{\mu}(u_.), s^2] = f[\hat{\mu}(u_.)] + \frac{As^2}{2\nu}[1-\lambda^2(u_.)] \quad . \tag{28}$$

For the simple square root transformation $A = 2$ and formula (28) coincides with (24).

One might regret that formula (26) involves the infinite series (27). However, in several examples conputed by the authors, this series happens to converge very rapidly so that a satisfactory approximation was reached by computing its first two or three terms.

The transformations that are currently most useful in meteorological work are the square root and the logarithmic transformations. They appear to work satisfactorily with precipitation and runoff data, one or the other having a slight advantage, depending upon the circumstances. However, with the growth of experimentation on weather control, certain other transformations are likely to become useful. For example, in the evaluation of experiments with lightning and/or hail prevention the arcsine and the archyperbolic-sine transformations may prove efficient in transforming the frequencies of the relevant events so as to satisfy the conditions of normal tests. For these reasons, we apply the general formulae to deduce the unbiased estimate $\theta[\hat{\mu}(u_.), s^2]$ appropriate to each of the several cases. For the sake of completeness, we reproduce first the formulae for $\hat{\mu}(u_.)$, $\lambda^2(u_.)$ and s^2,

for the two most important cases of s = 1 and s = 2. These are known formulae, from the theory of least squares.

6. SOME FORMULAE FROM THE THEORY OF LEAST SQUARES. Although the theory of least squares was developed some 150 years ago, particularly by Gauss [6], and although the relevant formulas are continuously being "developed" anew, we reproduce them here for the sake of completeness of the present paper. Particularly, we consider the two most important cases where the evaluation of the experiment is based on s = 1 and on s = 2 control areas. We deal with the method of estimating the transformed target precipitation expected to fall from a non-seeded storm when the transformed control precipitation has preassigned values: $u_{\cdot} = u_1$ in the case s = 1 and $u_{\cdot} = (u_1, u_2)$ in the case s = 2. Thus, all the operations indicated are to be performed on the data $(u_{\cdot j}, v_j)$ referring to the observed non-seeded storms. If we decide to accept assumption (iii) and it is necessary to calculate a similar estimate for seeded storms, then the calculations indicated by the formulas will have to be performed on data $(*u_{\cdot j}, *v_j)$, referring to the observed seeded storms.

All of the relevant formulae are expressed most conveniently in terms of sample means and of sample variances and covariances. Given some n numbers $a_1, a_2, \ldots, a_n$, the symbol $\bar{a}$ will denote their arithmetic mean and the symbol s_a^2 their sample variance,

$$(29) \qquad s_a^2 = \frac{1}{n}\sum (a_j - \bar{a})^2 = \frac{1}{n^2}\left[n\sum a_j^2 - \left(\sum a_j\right)^2\right] .$$

where all the summations extend over j = 1,2,...,n. For a given group of n pairs of numbers (a_j, b_j), the sample covariance is defined by

$$(30) \qquad s_{ab} = \frac{1}{n}\sum (a_j - \bar{a})(b_j - \bar{b}) = \frac{1}{n^2}\left[n\sum a_j b_j - \sum a_j \sum b_j\right| .$$

The sample variances of the transformed amounts of precipitation in the two controls will be denoted by s_1^2 and s_2^2, respectively. The sample variance of the transformed precipitation in the target will be denoted by s_v^2. The three covariances will be denoted s_{12}, s_{1v} and s_{2v}. Further, $\bar{u}_1$, $\bar{u}_2$ and $\bar{v}$ will indicate the observed mean amounts of rain, in transformed units.

We consider first the case s = 1 of one control area, say the first control area. If some other control area is the unique area, the formulae for it may be written by analogy. Suppose that the particular preassigned

transformed control precipitation, for which we want to estimate the expected precipitation in the target, is designated simply as u_1. Then we have

$$\hat{\mu}(u.) = \bar{v} + \frac{s_{1v}}{s_1^2}(u_1 - \bar{u}_1). \tag{31}$$

Further

$$\lambda^2(u.) = \frac{1}{n}\left[1 + \frac{(u_1 - \bar{u}_1)^2}{s_1^2}\right] . \tag{32}$$

It is clear that $\lambda^2(u.)$ is quadratic in u_1, attains its minimum of $1/n$ when $u_1 = \bar{u}_1$, and grows without limit as u_1 diverges from $\bar{u}_1$. Finally

$$s^2 = \sum\left[v_j - \bar{v} - \frac{s_{1v}}{s_1^2}(u._j - \bar{u}.)\right]^2 \tag{33}$$

$$= n\left[s_v^2 - \frac{s_{1v}^2}{s_1^2}\right] .$$

The corresponding number of degrees of freedom is $\nu = n-2$.

In the case of $s = 2$ control areas,

$$\hat{\mu}(u.) = \hat{\mu}(u_1, u_2) = \bar{v} + \alpha_1(u_1 - \bar{u}_1) + \alpha_2(u_2 - \bar{u}_2), \tag{34}$$

where

$$\alpha_1 = \frac{1}{\Delta}(s_2^2 s_{1v} - s_{12} s_{2v}) \tag{35}$$

$$\alpha_2 = \frac{1}{\Delta}(s_1^2 s_{2v} - s_{12} s_{1v}), \tag{36}$$

with

$$\Delta = s_1^2 s_2^2 - s_{12}^2. \tag{37}$$

Further,

$$(38)\qquad \lambda^2(u.) = \lambda^2(u_1,u_2) = \frac{1}{n}\left\{1 + \frac{1}{\Delta}\, s_2^2\left[(u_1-\bar{u}_1)^2 - 2s_{12}(u_1-\bar{u}_1)(u_2-\bar{u}_2) + s_1^2\,(u_2-\bar{u}_2)^2\right]\right\}$$

$$= \frac{1}{n}\left\{1 + \frac{(u_1-\bar{u}_1)^2}{s_1^2} + \frac{s_1^2}{\Delta}\left[(u_2-\bar{u}_2) - \frac{s_{12}}{s_1^2}\,(u_1-\bar{u}_1)\right]^2\right\}$$

The second line of formula (38) indicates that, in this case also, the minimum value $1/n$ of $\lambda^2(u.)$ is attained when $u_1 = \bar{u}_1$ and $u_2 = \bar{u}_2$. Further, as u_1 diverges from $\bar{u}_1$ and u_2 diverges from $\bar{u}_2$, the value of $\lambda^2(u.)$ grows indefinitely. Finally, the sum of squares of residuals is

$$(39)\qquad S^2 = n\left\{s_v^2 - \frac{1}{\Delta}\left[s_2^2\, s_{1v}^2 - 2s_{12}\, s_{1v}\, s_{2v} + s_1^2\, s_{2v}^2\right]\right\}.$$

Here, the number of degrees of freedom is $\nu = n-3$.

7. UNBIASED ESTIMATE OF EXPECTED PRECIPITATION IN THE TARGET WHEN THE SQUARE ROOT TRANSFORMATION IS USED. As already indicated, the minimum variance unbiased estimate $\hat{\theta}[\hat{\mu}(u.), S^2]$ of expected precipitation in the target without seeding, when the transformed precipitation in the control has values $u. = (u_1,u_2, \ldots, u_s)$ is given by formula (24). When the number of controls is either $s = 1$ or $s = 2$, the quantities $\hat{\mu}(u.)$, $\lambda^2(u.)$ and S^2 appearing in this formula are calculated from (31), (32) and (33) in case $s = 1$, and from (34) to (39) when $s = 2$. Only one further detail need be added.

If we want to evaluate the experiment under condition (iii) and to estimate the expectation of the extra precipitation ascribable to seeding that may have fallen in the target on the average, during all the m storms that were actually seeded, as indicated in formula (10), then, generally, this involves separate calculation of distinct estimates $\hat{\theta}[\hat{\mu}(*u_j), S^2]$ of $\theta(*u_j)$, where $j = 1,2,\ldots,m$, and then computing their arithmetic mean, say

$$(40)\qquad \hat{\theta}_0 = \frac{1}{m}\sum_{j=1}^{m} \hat{\theta}\left[\hat{\mu}(*u_j), S^2\right].$$

However, with the square root transformation, certain shortcuts are possible which might be worthwhile. In fact, using formula (24), we have

$$(41) \qquad \hat{\theta}_0 = \frac{1}{m}\sum_{j=1}^{m}\left[\hat{\mu}(*u_j)\right]^2 + \left[1 - \frac{1}{m}\sum_{j=1}^{m}\chi^2(*u_j)\right]\frac{s^2}{\nu} .$$

If there is just one control and, thus s = 1, then simple algebra reduces (41) to

$$(42) \quad \hat{\theta}_0 = \left[\bar{v} + \frac{s_{12}}{s_1^2}(*\bar{u}_1 - \bar{u}_1)\right]^2 + \frac{s_{12}^2 *s_1^2}{s_1^4} + \left\{1 - \frac{1}{n}\left[1 + \frac{*s_1^2 + (*\bar{u}_1 - \bar{u}_1)^2}{s_1^2}\right]\right\}\frac{s^2}{\nu} ,$$

where, as formerly, the asterisk on the left of a symbol indicates that the value of this symbol is to be calculated for the seeded storms. Thus, for example $*s^2$ means the sample variance of the m transformed precipitation amounts deposited by the seeded storms in the control area.

If s = 2, so that there are two control areas, then, by a similar procedure,

$$(43) \quad \theta_0 = \left[\bar{v} + \hat{\alpha}_1(*\bar{u}_1 - \bar{u}_1) + \hat{\alpha}_2(*\bar{u}_2 - \bar{u}_2)\right]^2 + \hat{\alpha}_1^2 *s_1^2 + 2\hat{\alpha}_1\hat{\alpha}_2 *s_{12}$$

$$+ \hat{\alpha}_2 *s_2 + \left\{1 - \frac{1}{n}\left[1 + \frac{1}{\Delta}\left[s_2^2\left(*s_1 + (*\bar{u}_1 - \bar{u})^2\right)\right.\right.\right.$$

$$\left.\left.\left. - 2s_{12}\left(*s_{12} + (*\bar{u}_1 - \bar{u}_1)(*\bar{u}_2 - \bar{u}_2)\right) + s_1^2\left(*s_2^2 + (*\bar{u}_2 - \bar{u}_2)^2\right)\right]\right]\right\}\frac{s^2}{\nu} ,$$

where $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the estimates (35) and (36) of the partial regression coefficients of V on U_1 and U_2, respectively.

Remark. It is interesting to notice that the formulae given here referring to the square root transformation do not require that the transformed variables be normally distributed. The only conditions of their validity is that the regression of the transformed target rain on the transformed control rain be linear and that the residual variance be constant.

8. UNBIASED ESTIMATE OF EXPECTED PRECIPITATION IN THE TARGET WHEN THE LOGARITHMIC TRANSFORMATION IS USED. If the normalizing transformation is logarithmic, then the transforming function is

$$(44) \qquad Y = f_0(V) = 10^V = e^{kV}$$

with k denoting the natural logarithm of 10. It is easy to see that this function belongs to the category we call recursive. In fact, differentiating (44) twice with respect to V we obtain

$$f_0''(V) = k^2 e^{kV} = k^2 f_0(V). \tag{45}$$

With reference to (25) it follows that in this case A = 0 and B = k^2. Thus, from (26), the minimum variance unbiased estimate of $\theta(u.)$ is given by

$$\hat{e}\left[\hat{\mu}(u.), s^2\right] = \phi\left\{\left[1 - \lambda^2(u.)\right] k^2 s^2, \nu\right\} 10^{\hat{\mu}(u.)} . \tag{46}$$

Unfortunately, in this case no shortcut exists similar to that available for the square root transformation and, if it is desired to estimate the average expected increase in rain from m seeded storms, formula (46) has to be computed for each of these storms separately. Unpleasantly, this will involve the evaluation of the series (27) m times, for m different values of the independent variable.

9. THE ANGULAR TRANSFORMATION OF FREQUENCIES. If an experiment is reduceable to the observation of several sets of units, each set containing the same number n of units and if the observations yield numbers Y of those units which are distinguished by the presence of certain specific characteristics, then a satisfactory "normalization" of the data is occasionally achieved by the so-called angular transformation

$$V = \arcsin\sqrt{\frac{Y+c}{n+2c}} . \tag{47}$$

Anscombe [7] suggests the value of c = 0.3 or 0.4. With reference to weather control experimentation, n may mean the number of seeded storms in a locality to be compared with an equal number of non-seeded storms. Y may mean the number of those storms in either group which are accompanied by hail.

The transforming function corresponding to (47) is

$$f_0(V) = (n+2c) \sin^2 V - c = \frac{1}{2}\left[n - (n+2c) \cos 2V\right]. \tag{48}$$

Differentiating (48) twice and performing easy transformations, it is found that

$$f_0''(V) = 2(n+2c) \cos 2V = 2n - 4f_0(V) \tag{49}$$

It follows that the transforming function (48) is of the recursive type, with A = 2n and B = -4. Consequently, the minimum variance

unbiased estimate of $\theta(u.)$ is given by the second general formula (26) which reduces to

$$(50)\qquad \hat{\theta}[\hat{\mu}(u.),s^2] = \phi\{-4[1-\lambda^2(u.)]s^2,\nu\}\{f[\hat{\mu}(u.)] - n/2\} + n/2$$

$$= \frac{1}{2}\left[n-(n+2c)\,\phi\{-4[1-\lambda^2(u.)]s^2,\nu\}\cos 2\hat{\mu}(u.)\right] .$$

10. HYPERBOLIC ARCSINE TRANSFORMATION. For integer valued variables X having a rather skew distribution, the hyperbolic arcsine transformation, first used by Beall, might be useful. In particular, this transformation might be applicable to the normalization of such variables as the number X of lightning strikes per storm. We will write it in the form

$$(51)\qquad V = \operatorname{arcsinh}\sqrt{X} \ .$$

The corresponding transforming function is

$$(52)\qquad X = f_0(V) = [\sinh V]^2 = \frac{1}{4}\left[e^{2V} + e^{-2V} - 2\right] .$$

Differentiating (52) twice and performing easy calculations, we find

$$(53)\qquad f_0''(V) = 2 + 4f_0(V).$$

Hence, (52) is a recursive function with A = 2 and B = 4 and, according to (26),

$$(54)\qquad \hat{\theta}[\hat{\mu}(u.),s^2] = \frac{1}{2}\left[\phi\{4[1-\lambda^2(u.)]s^2,\nu\}\cosh 2\hat{\mu}(u.) - 1\right] .$$

REFERENCES

1 A. M. Mood, Introduction to the Theory of Statistics, McGraw-Hill, New York, 1950.

2 H. C. S. Thom, "A statistical method of evaluating augmentation of precipitation by cloud seeding," Final Report of the Advisory Com. on Weather Control, Vol. 2 (1957), pp. 5-25.

3 J. Neyman and E. L. Scott, "Correction for bias introduced by a transformation of variables," Annals Math. Stat., Vol. 31 (1960), to appear.

4 E. L. Lehmann and H. Scheffé, "Completeness, similar regions and unbiased estimation, Part I," Sankhyā, Vol. 10 (1950), pp. 305-340.

5 J. W. Tukey, "On the comparative anatomy of transformations," Annals Math. Stat., Vol. 28 (1957), pp. 602-632.

6 C. F. Gauss, Abhandlunger zur Methode der kleinsten Quadrate, Berlin, 1887. Transl. from Latin by A. Borsch and P. Simon.

7 F. J. Anscombe, "The transformation of Poisson, binomial and negative-binomial data," Biometrika, Vol. 35 (1948), pp. 246-254.

MATHEMATICAL AND STATISTICAL PRINCIPLES UNDERLYING CHEMICAL CORPS INSPECTION PROCEDURES FOR PRODUCT VERIFICATION

Henry Ellner
U.S. Army Chemical Center and Chemical Corps
Materiel Command, Directorate for Quality Assurance

1. INTRODUCTION. In accordance with Standardization Manual M205, issued 9 April 1958 by DOD, each military specification is required to include in the beginning of the quality assurance provisions, section 4, the following statement:

> "Unless otherwise specified herein the supplier is responsible for the performance of all inspection requirements prior to submission for Government inspection and acceptance. Except as otherwise specified, the supplier may utilize his own facilities or any commercial laboratory acceptable to the Government. Inspection records of the examinations and tests shall be kept complete and available to the Government as specified in the contract or order."

The contractor is thus obliged to assure himself that supplies he offers for acceptance conform with contractual requirements. The Army inspector, under AR 715-20* is enjoined to verify the inspection performed by the contractor and to establish the reliability of the contractor's inspection records prior to acceptance of the submitted supplies. Limited product inspection by the Government inspector and review of the supplier's inspection system are the means prescribed for authenticating the contractor's inspection data.

As stated by Standardization Manual M205, sampling is an important factor in determining compliance with requirements. While details of sampling will vary with the commidities, where applicable, MIL-STD-105 (Sampling Procedures and Tables for Inspection by Attributes) is referenced in specifications. In conjunction with MIL-STD-105, the quality assurance provisions of the specification include one or more classifications of defects. Now using the acceptable quality level (AQL) in the specification as an index to the plans of MIL-STD-105, the contractor is obligated to sample and inspect as prescribed. Sampling inspection is also the modus operandi of the Government inspector for verifying the inspection data recorded by the contractor.

2. VERIFICATION BY SAMPLING. The Government cannot expect its verification data to duplicate the contractor's inspection results exactly since the number of defectives in each sample is a variate dependent upon the true (but unknown) percent defectives in the lot sampled. Chance can, therefore, be responsible for the difference in the proportions of defectives observed in any two samples compared. Wide discrepancies in results may be due to non-random (biased) samples or to failure to recognize a defective as spelled out in the classification of defects. The problem is to set up criteria so that disparities arising by chance alone are differentiated, considering costs and consequences, from disparities arising from improper inspection practices. Furthermore, any general

---

* Policy background pertaining to product verification inspection is discussed in Appendix 1. This Appendix is under the authorship of Joseph Mandelson.

tendency for significant discrepancies to arise must be recognized by frequent testing so that appropriate action can be taken to safeguard Government interests.

When the problem is as stated above the purpose and procedures for accomplishing verification sampling are conceptually simplified. A decision whether to accept or reject a lot, in accordance with acceptance criteria of a sampling plan of the single, double or multiple type selected from MIL-STD-105, is not involved. The decision as to reliability of contractor inspection results is distinct from the decision to accept or reject a lot, although the latter decision may be contingent upon the former. Verification sampling inspection then has as its primary purpose the establishment of the validity of the contractor's inspection data by checking his sampling results against independent sampling inspections. The size of the sample required and the frequency of performing verification sampling depend upon the power of the test to detect significant differences between the paired samplings and the establishment of an objective degree of rational belief in the existence of a state of statistical control over the fluctuating differences observed. This belief can be bolstered by evaluation of the contractor's inspection system and by independent assessment of the qualtiy of product offered to the Government for acceptance.

3. HOMOGENEITY OF ATTRIBUTE SAMPLING DATA. As defined in Supply and Logistics Handbook H105, in inspection by attributes the unit of product is classified simply as defective or nondefective with respect to a given requirement or a set of requirements. The requirement may be an individual checkpoint and the set may be a group of characteristics of equal importance listed under a single AQL in the specification. In the following development, we shall assume that, even when measurement along a contunious numerical scale is possible, such measurement will be classified as conforming or non-conforming with specification limits.

Let us now suppose that the contractor has drawn a random sample in accordance with MIL-STD-105 from an inspection lot and has noted the number of conforming and non-conforming items in the initial sample. The Government inspector has proceeded likewise by selecting an independent sample from the same lot. In the analyses that follow, we shall asuume that the lot size is large relative to the total sample size (say, at least 8:1); or, if the lot size is proportionately small, that the samples are drawn without replacement until a comparison has been made. The results of inspection are denoted symbolically in a 2 X 2 table as below:

TABLE 1.

Notation for 2 X 2 Contingency Table

	Defective	Nondefective	Total
Contractor's Sample	d_c	$n_c - d_c$	n_c
Government's Sample	d_g	$n_g - d_g$	n_g
Total	d_t	$n_t - d_t$	n_t

The data are recorded to decide whether the results of inspecting two samples, one the size n_c and the other the size n_g, which are found to contain d_c and d_g defectives, respectively, are significantly different. A common test of significance, for attribute data classified in two ways as shown, is the chi-square test [13] and equivalent alternates. When the expected number of defectives is small, say less than five, Fisher's exact test ([9], Section 21.02) is generally advised. For routine testing these techniques all involve extensive computation, and consequently are not suitable for verification purposes. Short cut procedures [8, 11, 12, 16] devised to meet this problem, including nomograms and extensive tabulations of Fisher's "exact" test, are likewise wanting in that multiple entries are necessary or that tables required are too lengthy and numerous.

A test for homogeneity, applicable when the overall proportion of defectives d_t/n_t is small, say 0.20 or less, is one which compares samples from populations known to give the Poisson type of distribution. Przyborowski & Wilenski [15] considered two observations (in our notation: d_c and d_g) originating from two Poisson-distributed populations with unknown means, and for the symmetrical case $n_c = n_g$ they proposed an "exact" test for the equality of these means. Barnard [1] extended their method to the case $n_c \neq n_g$ reducing the procedure to a simple test for the variance - ratio F. Bross and Kasten [5] derived a related technique for the case $n_c \neq n_g$ and published charts for avoiding or greatly reducing computations for the analysis of fourfold contingency tables. What was apparently a very different test from Barnard's was proposed by Cox [6], but their similarity has been shown by Barton [2]. However, Cox's method has certain advantages over Barnard's which make it preferable for use in product verification.

4. "EXACT" TESTS OF SAMPLES FROM TWO POISSON SERIES. Before the advantages of Cox's method can be discussed it will be necessary to derive the "exact" test for comparing two Poisson-distributed observations. Suppose d_c and d_g of Table 1 approximately follow independent Poisson distributions so that:

$$(1) \quad P(d_c, d_g \mid p'_c, p'_g) = P(d_c) \cdot P(d_g) = \frac{e^{-p'_c n_c}(p'_c n_c)^{d_c}}{d_c!} \cdot \frac{e^{-p'_g n_g}(p'_g n_g)^{d_g}}{d_g!},$$

where:

p'_c = the expected fraction defective in the contractor's sample n_c

p'_g = the expected fraction defective in the Government inspector's sample n_g.

Under the null hypothesis $p'_c = p'_g = p'_o$ so that Equation (1) reduces to:

$$(2) \quad P(d_c,d_g \mid p'_o) = \frac{e^{-p'_o(n_c+n_g)} (p'_o)^{d_t} n_c^{d_c} n_g^{d_g}}{d_c!\ d_g!} ,$$

which can be rewritten as:

$$(3) \quad P(d_c,d_g \mid p'_o) = P(d_g \mid d_t)\ P(d_t \mid p'_o)$$

$$= \frac{d_t!}{d_c!\ d_g!} \cdot \frac{n_c^{d_c} n_g^{d_g}}{(n_c+n_g)^{d_c}(n_c+n_g)^{d_g}} \cdot \frac{e^{-p'_o(n_c+n_g)} (p'_o n_c + p'_o n_g)^{d_t}}{d_t!}$$

But we need the probability of getting some pair of results having the same total $d_c + d_g = d_t$; and so the relative probability, on the null hypothesis, of getting the pair (d_c,d_g) out of all results with the same total d_t is:

$$(4) \quad P(d_g \mid d_t) = \frac{P(d_g \mid d_t)\ P(d_t \mid p'_o)}{P(d_t \mid p'_o)}$$

$$= \frac{d_t!}{d_c!\ d_g!} \left(\frac{n_g}{n_c + n_g}\right)^{d_g} \left(\frac{n_c}{n_c + n_g}\right)^{d_c}$$

If we let $r = \frac{n_c}{n_g}$ then:

$$(5) \quad P(d_g \mid d_t) = \frac{d_t!}{d_c!\ d_g!} \left(\frac{1}{1 + r}\right)^{d_g} \left(\frac{r}{1 + r}\right)^{d_c}$$

We note that conditionally on d_t, d_g is binomially distributed with parameters, $\frac{1}{1 + r}$ and d_t, which can be used as the basis for a significance test. Accordingly:

$$(6) \quad F(y) = \sum_{y=d_g}^{d_t} \binom{d_t}{y} \left(\frac{1}{1 + r}\right)^{y} \left(\frac{r}{1 + r}\right)^{d_t - y} = I_{\frac{1}{1 + r}} (d_g, d_c + 1) ,$$

where $I_x(p,q)$ is the incomplete β - function representation of a sum of binomial probabilities.

If the only admissible alternative to the null hypothesis $p'_c = p'_g = p'_o$ is $p'_g > p'_c$ then the appropriate critical region, in the Neyman-Pearson sense, for rejection of the null hypothesis is defined by $d_c \leq k_1 (d_t, \alpha)$ or $d_g \geq k_2 (d_t, \alpha)$,

where α is the risk of the first kind of error and where

$$(7) \quad P\left\{d_g \geq K_2 (d_t, \alpha) \mid d_t, p'_c = p'_g\right\} \leq \alpha .$$

For the "exact" test this may be expressed by:

$$(8) \quad I_{\frac{1}{1+r}} (d_g, d_c + 1) \leq \alpha .$$

This inequaltiy may be written in terms of the probability distribution function $P_{f_1,f_2}(F)$ of the F distribution with (f_1,f_2) degrees of freedom since:

$$P_{f_1,f_2}(F) = I_x (p,q)$$

where $f_1 = 2q$, $f_2 = 2p$ and $F = \frac{p}{q} \frac{1-x}{x}$ with the result that

$$(9) \quad P_{2d_c+2,2d_g}\left(\frac{r\, d_g}{d_c + 1}\right) \leq \alpha .$$

Inequaltities(8) and (9) establish a level of significance which does not exceed α. The true level of significance depends upon the unknown p'_o and may in some cases for small (d_c,d_g) be considerably less than α.

5. COX'S "APPROXIMATE" TESTS FOR POISSON VARIATES. In inverse Poisson sampling, with d fixed, the number of sample items n drawn in sequence up to the d th event is distributed as $(2p')^{-1}\chi^2_{2d}$, where χ^2_{2d} denotes a chi-square variate with 2d degrees of freedom and p' represents the true rate. For direct Poisson sampling in which the number of events d occurring in a fixed n is observed, we have

$$(10) \quad P(x \geq d) = \sum_{x=d}^{\infty} \frac{e^{-p'n}(p'n)^x}{x!} = P\left(\frac{1}{2p'}\chi^2_{2d} \leq n\right), \text{ and}$$

$$(11) \quad P(x \geq d+1) = P\left(\frac{1}{2p'}\chi^2_{2d+2} \leq n\right) .$$

Cox suggested an approximation to P (x>d) in which d is treated as a continuous variate by taking a quantity intermediate between (10) and (11):

$$(12) \quad P(x>d) \simeq P\left(\frac{1}{2p'} \chi^2_{2d+1} \leq n\right),$$

which implies that probabilities are calculated as if

$$(13) \quad 2\,p'n \text{ is distributed as } \chi^2_{2d+1}$$

When two populations with fraction defectives p'_c, p'_g are compared by means of samples n_c, n_g which exhibit d_c, d_g defectives, then, from (12) we compute the ratio:

$$(14) \quad \frac{2p'_c n_c}{2d_c+1} \div \frac{2p'_g n_g}{2d_g+1}$$

which is distributed approximately as F with $(2d_c+1, 2d_g+1)$ degrees of freedom. Thus, we may test the hypothesis that $p'_c = p'_g = p'_o$ against the alternate hypothesis that $p'_g > p'_c$ by referring

$$(15) \quad F = r\,\frac{(d_g+0.5)}{d_c+0.5}$$

to the F tables with $(2d_c+1, 2d_g+1)$ degrees of freedom for the appropriate α percent point.

This may be represented by

$$(16) \quad P_{2d_c+1, 2d_g+1}\left(r\,\frac{d_g+0.5}{d_c+0.5}\right) \leq \alpha$$

or

$$(17) \quad I_{\frac{1}{1+r}}(d_g+0.5,\ d_c+0.5) \leq \alpha$$

It is now clear that the "exact" tests given by (8) and (9) have been modified slightly to yield the approximate tests of (16) and (17). The modification has the effect of making the true level of significance less dependent upon the unknown p'_o and to approximate the nominal value of α when averaged over d_t.

6. POWER FUNCTION OF TESTS FOR POISSON VARIATES. The Neyman-Pearson theory of tests considers all tests of the same size and lays down objective standards for selecting the best test. The theory introduces the term, "power of a test," relative to the alternate hypothesis, to denote the probability of correctly rejecting the null hypothesis when an alternative is true. Of all tests at a given significance level, the most preferred is the one which has the maximum power relative to all the alternate hypothesis considered. The probability of rejecting the null hypothesis H_o, regarded as a function of H', where H' is any of the admissible alternates to H_o, is called the power function of the test. If we commence with the determination of the critical region subject to (7) we can calculate the power function of a given test of significance. Thus, for the "exact" test all points satisfying (8) or (9) are entered in (1) and the absolute probabilities are summed. Similarly, for the "approximate" test all points satisfying (16) or (17) are entered in (1) for addition of the absolute probabilities. Tables 2 and 3 provide the actual probabilities associated with the respective tests for a one-sided test of the null hypothesis $p'_g = p'_c$ against the alternatives $p'_g = 3p'_c$ and $p'_g = 4.5\ p'_c$ for $r = 1, 2, 3, 5$ and 8, respectively, over a range of nuisance parameters, $p'_c n_c$, which may be encountered in practice.

The arrangement of Tables 2 and 3 clearly reveals that the significance level α is a function of the expected number of defectives in the contractor's sample and the ratio of the contractor's sample size to the size of the Government's verification sample. For the "exact" test, under the null hypothesis, $p'_g = p'_c$, the quantity α increases about tenfold on the average as $p'_c n_c$ increases from 0.75 to 9.00. In contrast, for the "approximate" test, α increases only 1.5 times on the average over the same range of $p'_c n_c$. Furthermore, the average level of significance of the thirty entries summed over the five tabular values of r for the "exact" and "approximate" tests are 0.015 and 0.052, respectively. The conclusion is that the "approximate" test more effectively controls the size of the test at the significance level of 0.05 than the "exact" test.

Since we can generally estimate $p'_c n_c$ from the contractor's record of inspection results and the AQL under which he is operating, we can select the power of test by adjusting the sample size ratio r commensurate with relative fraction defective, p'_g/p'_c, which should be detected if it exists. This power can be further augmented by simple pooling of inspection results for a given r until the expected number of defectives for the contractor's samples exceeds the desired value of $p'_c n_c$. Birnbaum [4] has considered various methods of comparing two Poisson processes in terms of the ratio of their parameters, and suggests for fixed samples an accumulation of observations until the total number of defectives d_t is sufficient to yield the power of test desired.

7. COMBINATION OF TESTS OF POISSON VARIATES. When the sample size ratio r is varied or the class of defects considered is not maintained constant so that pooling of inspection results from a sequence of lots is inappropriate for the methods represented by (8) or (9) and (16) or

TABLE 2

Power of Extended P-W* 'Exact' Test at Nominal Significance Level 0.05 for Hypothesis $p'_g = p'_c$ Against Alternatives $p'_g > p'_c$

	r = 1			r = 2			r = 3			r = 5			r = 8		
	p'_g/p'_c			p'_g/p'_c			p'_g/p'_c			p'_g/p'_c			p'_g/p'_c		
$p'_c n_c$	1	3	4.5	1	3	4.5	1	3	4.5	1	3	4.5	1	3	4.5
.75	.001	.040	.139	.003	.062	.144	.001	.023	.065	.005	.042	.082	.003	.026	.059
1.50	.004	.173	.433	.011	.177	.372	.004	.084	.217	.011	.093	.180	.009	.070	.136
2.25	.010			.020			.008			.014			.013		
3.00	.015	.428	.790	.024	.351	.650	.013	.244	.513	.017	.190	.372	.015	.135	.280
4.50	.022			.029			.020			.020			.020		
9.00	.032			.031			.027			.026			.028		

*Przyborowski, J. & Wilenski, H. [15]

TABLE 3

Power of Cox's* 'Approximate' Test at Nominal Significance Level 0.05 for Hypothesis $p'_g = p'_c$ Against Alternatives $p'_g > p'_c$

	r = 1			r = 2			r = 3			r = 5			r = 8		
	p'_g/p'_c			p'_g/p'_c			p'_g/p'_c			p'_g/p'_c			p'_g/p'_c		
$p'_c n_c$	1	3	4.5	1	3	4.5	1	3	4.5	1	3	4.5	1	3	4.5
.75	.020	.214	.407	.028	.188	.352	.014	.102	.202	.070	.200	.284	.044	.144	.198
1.50	.049	.389	.652	.054	.336	.554	.029	.220	.409	.072	.241	.356	.049	.184	.275
2.25	.062	.488	.774	.066	.429	.678	.040	.318	.550	.062	.264	.432	.050	.213	.354
3.00	.065	.564	.862	.070	.495	.761	.048	.390	.652	.056	.315	.515	.051	.255	.424
4.50	.060	.702	.955	.068	.591	.869	.055	.500	.790	.053	.394	.653	.053	.315	.535
9.00	.049	.928	.999	.053	.810	.986	.056	.743	.959	.049	.593	.865	.053	.466	.754

* Cox, D. R. [6]

(17) an omnibus type of test is required. This test can serve to combine all of the evidence obtained by means of verification sampling to provide a single measure of confidence in the contractor's inspection results.

From the α risks associated with the "exact" and "approximate" tests under the null hypothesis we can expect a certain frequency of significant differences. Further, from the β risks associated with these tests we can expect a certain frequency of erroneous acceptances of false hypotheses. Accordingly, it is not correct to reject or accept the general hypothesis that the contractor's inspection data are as a whole unreliable as a consequence of the individual lot comparisons, which taken separately appear to yield either significant or non-significant results. The over-all test calls, therefore, for the combination of a number of independent tests of significance. Fisher ([9] , Section 21.1) has given a general method for combining the probabilities of several mutually independent tests. A number of other writers have discussed and illustrated this problem, but Birnbaum [3] has shown that Fisher's method is to be preferred for its somewhat more uniform sensitivity to the alternatives of interest.

The over-all test developed by Fisher deals with continuous variables. It will yield biased results if applied directly to probabilities derived from the "exact" test for Poisson variates. Lancaster [10], David and Johnson [7], Tocher [17] and Pearson [14] have considered the difficulties encountered by the combination of tests based on discontinuous variates. Since Cox's "approximate" test treats the number of events, d_c, d_g as continuous variates the probabilities obtained can be handled on a practical basis by application of Fisher's probability integral, which may be defined generally as follows:

Let $p(\chi)$ be the probability density function of a continuous random variable χ in the interval $a \leq \chi \leq b$, where $p(\chi) = 0$ for $\chi < a$ or $\chi > b$. Then if

$$P = \int_a^{\chi} p(\chi)\, d\chi , \tag{18}$$

P is uniformly distributed in the interval (0,1) and $x = -2 \log_e P$ is distributed as χ^2 with 2 degrees of freedom.

If now we combine k independent probabilities, the combined probability is the product of the k separate probabilities, or

$$\begin{aligned} \sum(z_i) &= -2 \log_e (P_1 P_2 \ldots P_k) \\ &= -2 \sum_{i=1}^{k} \log_e P_i , \end{aligned} \tag{19}$$

and so has the χ^2 distribution with 2 k degrees of freedom. Thus, by means of the probability integral transformation, any number of probabilities $P_1, P_2, \ldots, P_k$ may be converted to a χ^2 value and, using the additive

properties of the χ^2 distribution, may be summed together with the degrees of freedom to yield from published tables an over-all probability. The application of these results to continuous populations is straightforward.

For discrete populations, such as the binomial represented by (5), the over-all probability is biased when the null hypothesis is true. The expectation of χ^2 for discontinuous variates is always below the theoretical value of 2. Thus, for the case $d_g+d_c = 4$ and $r = 1$ we obtain, under the null hypothesis, the binomial $(1/2 + 1/2)^4$ and find from Table 4 below for a one-sided comparison that the expectation of $-2 \log_e P_i$ is 1.241 and the variance of the distribution is 3.527.

TABLE 4

Distribution of Probability Integral Transformation Applied to "Exact" Test for Case of Binomial $(1/2 + 1/2)^4$

No. of Events		Relative Frequency	Cumulative Probability	Probability Integral Transform
d_c	d_g	of d_c, d_g	P_i	$-2 \log_e P_i$
4	0	0.0625	1.0000	0
3	1	0.2500	0.9375	0.1291
2	2	0.3750	0.6875	0.7494
1	3	0.2500	0.3125	2.3263
0	4	0.0625	0.0625	5.5452

	Expectation	Variance
χ^2 with 2 D.F. (theoretical)	2.000	4.000
$-2 \log_e P_i$	1.241	3.527

Similarly, for the case of the binomial $(1/3 + 2/3)^5$ which can be derived from (5) the expectation of χ^2 is 1.314 and the variance of the distribution is 2.482. In contrast, Cox's "approximate" method for the same distribution as shown in Table 5 below yields a χ^2 expectation of 2.042 and a variance of 4.393.

TABLE 5

Distribution of Probability Integral Transformation Applied to "Exact" and "Approximate" Tests for Case of Binomial $(1/3+2/3)^5$

No. of Events		Relative Frequency	Probability Integral Transforms for Probabilities Derived from	
d_c	d_g	of d_c, d_g	"Exact" Test	"Approximate" Test
5	0	0.131687	0.0000	0.0796
4	1	0.329218	0.2824	0.6570
3	2	0.329218	1.2357	2.0488
2	3	0.164609	3.1225	4.4886
1	4	0.041153	6.1903	8.3082
0	5	0.004115	10.9862	14.6404

	Expectation	Variance
χ^2 with 2 D.F. (Theoretical)	2.000	4.000
"Exact" Test Probability Integral Transformation	1.314	2.482
"Approximate" Test Probability Integral Transformation	2.042	4.393

There is clearly considerable bias when the probability integral transformation is applied to the probabilities derived from the "exact" test. In contrast, Table 6 below indicates comparative lack of bias in the behavior of the "approximate" test when we wish to combine its results for a series of independent determinations to verify a common hypothesis, i.e., that the contractor's inspection records are reliable. The numerical results of Table 6 show that even for an extremely small number of observed defectives the continuity correction of the "approximate" test is very effective.

TABLE 6

Expectances and Variances of Binomially-Distributed Probability Integral Transformations Derived from "Approximate" Tests of Poisson Variates (one-sided comparison)

n \ p	$\frac{1}{1+r} = 1/2$		$\frac{1}{1+r} = 1/3$		$\frac{1}{1+r} = 1/4$		$\frac{1}{1+r} = 1/6$		$\frac{1}{1+r} = 1/9$	
d_t	E(z)	Var(z)	E(z)	Var(z)	E(z)	Var(z)	E(z)	Var(z)	E(z)	Var(z)
5	2.045	4.364	2.042	4.393	2.044	4.392	2.056	4.409	2.084	4.391
4	2.050	4.316	2.051	4.253	2.056	4.485	2.072	4.485	2.111	4.453
3	2.045	4.108	2.067	4.540	2.074	4.604	2.101	4.585	2.159	4.431
2	2.024	3.540	2.086	4.463	2.106	4.686	2.158	4.678	2.257	4.458
1	1.905	2.259	2.084	3.630	2.170	4.165	2.302	4.377	2.474	4.146

NOTES: (contd. from Table 6)

(a) $z = -2 \log_e I_{\frac{1}{1+r}} (d_g+0.5 \ , \ d_c+0.5)$

(b) z is disbributed as $\left(\frac{1}{1+r} + \frac{r}{1+r}\right)^{d_t}$ where $d_t = d_g + d_c$;

viz. $(p + q)^n$

(c) $E(\chi^2)_{D.F.=2} = 2.000$,

$\text{Var } (\chi^2)_{D.F.=2} = 4.000$.

8. TABLES FOR ACCOMPLISHING VERIFICATION INSPECTION. Data recorded as shown in Table 1 can be conveniently tested for statistical significance by means of a table providing critical limits.

For a given number of total defectives, d_t, observed in both the contractor's and Government inspector's samples, limits can be set for either d_g of d_c as indicated by (17) for a specified α. This arrangement enumerates the boundary points of the critical region of the test of significance. However, the Government inspector is more concerned with comparing his sample results, for a given d_c recorded by the contractor, against an "allowable number." Accordingly, the critical value for d_g, designated as $d_g(A)$ can be obtained from (17) for a specified r. When the critical number $d_g(A)$ is reached or exceeded, the Government inspector adopts a course of action on the premise that a discrepancy actually exists in the contractor's inspection system.

Critical limits for indicating a discrepancy in paired attribute sampling inspections are presented in Tables IA through IE* of Section III of the Chemical Corps Verification Handbook [19]. The five sections, A through E, correspond to sample size ratios of 1, 2, 3, 5 and 8, respectively. Two standards of significance were set: $\alpha \leq 0.05$ for $d_g(A)$ and $\alpha \approx 0.10$ for a "warning" limit $d_g(W)$. When the "warning" limit is reached, the Government inspector is alerted to look for a possible discrepancy in the contractor's inspection system.

The probability integral transformation,

(20) $$z = -2 \log_e I_{\frac{1}{1+r}} (d_g+0.5 \ , \ d_c+0.5)$$

for a given d_g, d_c at a specified r can be readily derived from Tables of the Incomplete Beta- Function [18] and natural logarithm tables, and tabulated for comparison against critical values of χ^2 in accordance with (19). To simplify the procedure for the Government inspector, a Table II,** "Check Ratings for Paired Attribute Sampling Insepction" and a Table III,**

---

* Table IE is illustrated in Appendix 2.

** Portions of Table IIE and Table III are illustrated in App. 2.

"Upper Critical Limits for Cumulative Check Ratings" have been included in Section III of the Chemical Corps Verification Handbook, Table II, which is subdivided into five sections corresponding to r = 1, 2, 3, 5, and 8, yields directly for a pair of values, d_c, d_g, the quantity 1/2 z. Table III is an extended table of the percentage points of the χ^2 distribution for even-numbered degrees of freedom. As Table III is used in conjunction with Table II, the critical values tabulated are $1/2\chi^2_{2k}$ for 2 k degrees of freedom, where k is the number of probabilities to be combined, i.e., number of lots verified. The warning and action limits in Table III have been set at the 0.10 and 0.01 significance levels, respectively, and the median value at the 0.50 level.

The accumulation of check ratings serves to summarize all available information concerning the reliability of the contractor's inspection results. Furthermore, the ratings establish an objective degree of confidence, in the existence of statistical control over the contractor's inspection practice compared with the Government's standards. Visual representation of the check ratings on semi-logarithmic graph paper, with 1/2 z plotted on the log scale and critical limits of $1/2\chi^2_{D.F.=2}$ imposed will be found useful for recording serially a common set of tests of significance.

9. ESTIMATING PRODUCT QUALITY. Product verification sampling has as its primary purpose the checking of the supplier's inspection records. However, the verification sampling results are also useful in providing an independent estimate of the contractor's "process average" and in furnishing an unbiased estimate of the quality of the conforming lots offered by the contractor for Government acceptance. Since the contractor's process average determines whether reduced, normal, or tightened inspection should be used, its validity should be established. The tolerance limits in MIL-STD-105 for a specified AQL can also be applied to the results of verification sampling. Normally, these limits should be applied to the process average derived from the results of the non-conforming lots as well as the conforming lots, since the process average reflects the average quality of product on which the supplier performs inspection.

The supplier's inspection results serve to segregate his inspection lots into conforming and non-conforming segments, and to determine the average percentage of defective items in the product represented by the samples inspected. Only when all lots are in conformance with acceptance criteria or when the product is manufactured under statistically controlled conditions can the process average computed by the contractor be used to furnish an unbiased estimate of the quality of product offered for Government acceptance.

Consider the O-C curve of a single sampling plan:

$$(21) \qquad Lp' = \sum_{0}^{c} \binom{n}{c} q'^{n-c} p'^{c}$$

where Lp' denotes the probability of acceptance of lots binomially controlled at quality p'. If the plan n,c is designed or selected so that $0<L_p<1$ then a portion of all lots from the controlled process will yield samples in conformance with the acceptance criterion, c, and the rest of the lots will yield samples which are not in conformance. For the conforming fraction, the number of defectives, d_c, in each accepted sample will vary from 0 to c, and for the non-conforming fraction d_c will vary from c + 1 to n. Since n is fixed for all lots submitted for inspection the mean number of defectives in the samples from conforming lots will be less than the mean number of defectives in samples from the non-conforming lots. The apparent difference in the estimated quality between the two fractions of lots submitted for inspection against the acceptance plan (n,c,) contradicts the original premise that the production of all lots was binomially controlled at a fixed p'. Accordingly, it is evident that sample results used to segregate lots cannot furnish an unbiased estimate of the respective fractions. This argument can be extended to the common case by using inverse probability for lots productd from different binomially controlled processes to demonstrate that acceptance sampling results cannot furnish unbiased estimates of the quality of the conforming segment offered for Government acceptance.

For an unbiased quality estimate, the following generalization can be used to obtain the best linear estimate of the percent defective p' of any lotted portion of product sampled independently by the Government inspector:

$$(22)\quad \hat{p} = \frac{N_1\hat{p}_1 + N_2\hat{p}_2 + \ldots + N_k\hat{p}_k}{N_1 + N_2 + \ldots + N_k},$$

where $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$ are the respective estimates of lot quality derived from Government sampling results of inspection lots with lot sizes $N_1, N_2, \ldots, N_k$.

If lot sizes are approximately equal the following estimate is an unbiased estimate of p':

$$(23)\quad \hat{p} = \frac{\sum_{k=1}^{k} (d_g)i}{\sum_{i=1}^{k} (n_g)i},$$

where (d_g) i denotes the number of defectives found in a sample of size $(n_g)i$ drawn from the ith lot checked by the Government.

When skip lot sampling is practiced by the Govermment inspector the lots selected for verification should be randomly selected to assure an unbiased estimate of p'. Usually skip lot sampling will not be as efficient as proportional sampling from each lot for estimating p'.

REFERENCES

[1] Barnard, G. A., "A test for homogeneity of Poisson series," Advisory Service on Statistical Method and Quality Control, Technical Report No. Q.C./R/18, Ministry of Supply, London (1944).

[2] Barton, D. E., "On the equivalence of two tests of equality of rate of occurrence in two series of events occurring randomly in time," Biometrika, Vol. 45 (1958), pp. 267-268.

[3] Birnbaum, Allan, "Combining independence tests of significance," Journal of the American Statistical Association, Vol. 49 (1954), pp. 559-574.

[4] Birnbaum, Allan, "Statistical methods for Poisson processes and exponential populations," Journal of the American Statistical Association, Vol. 49 (1954), pp. 254-266.

[5] Bross, Irwin D. J., and Kasten, E. L., "Rapid analysis of 2x2 tables," Journal of the American Statistical Association, Vol. 52 (1957), pp. 18-28.

[6] Cox, D. R., "Some simple approximate tests for Poisson Variates," Biometrika, Vol. 40 (1953), pp. 354-360.

[7] David, F. N., and Johnson, N. L., "The probability integral when the variable is discontinuous," Biometrika, Vol. 37 (1950), pp. 42-49.

[8] Finney, D. J., "The Fisher-Yates test of significance in 2x2 contingency tables," Biometrika, Vol. 35 (1948), pp. 145-156.

[9] Fisher, R. A. Statistical Methods for Research Workers, 8th Ed., Oliver & Boyd, Ltd., Edinburgh and London (1941).

[10] Lancaster, H. O., "The combination of probabilities arising from data in discrete distributions," Biometrika, Vol. 36 (1949), pp. 370-382.

[11] Latscha, R., "Tests of significance in a 2x2 contingency table: extension of Finney's table," Biometrika, Vol. 40 (1953), pp. 74-86.

[12] Mainland, Donald and Murray, I. M., "Tables for use in fourfold contingency tests," Science, Vol. 116 (1952), pp. 591-594.

[13] Paulson, Edward and Wallis, W. A., "Planning and analyzing experiments for comparing two percentages," Techniques of Statistical Analysis, Statistical Research Group Columbia University, McGraw-Hill Book Co., Inc., New York (1947).

[14] Pearson, E. S., "On questions raised by the combination of tests based on discontinuous distributions," Biometrika, Vol. 37 (1950), pp. 383-398.

[15] Przyborowski, J. and Wilenski, H., "Homogeneity of results in testing samples from Poisson series," Biometrika, Vol. 31 (1940), pp. 313-323.

[16] Swaroop, Satya, "Exact significance of difference in responses under two treatments," Indian Medical Research Memoirs, Memoir No. 35.

[17] Tocher, K. D., "Extension of the Neyman-Pearson theory of tests to discontinuous variates," Biometrika, Vol. 37 (1950), pp. 130-144.

[18] Tables of the Incomplete Beta-Function, edited by Karl Pearson, Biometrika Office, University College, London (1934).

[19] Tables for Accomplishing Product Verification Sampling Inspection, U. S. Army Chemical Corps Verification Inspection Handbook, Section III, prepared by Headquarters, U. S. Army Chemical Center and Materiel Command, Directorate for Quality Assurance.

APPENDIX 1.

Basis for Preparation of Section III, "Statistical Sampling and Assessment" of the Chemical Corps Verification Inspection Handbook

Joseph Mandelson

U. S. Army Chemical Center and Chemical Corps
Materiel Command, Directorate for Quality Assurance

1. References:

a. AR 715-20 dated 3 September 1957

b. Change 1, AR 715-20, dated 2 January 1958

2. Section III of the Chemical Corps Verification Inspection Handbook is intended to implement one important phase of Army inspection policy as established by References 1a and 1b. At this time it provides statistical tools for accomplishing the objectives of product verification inspection by attributes.

3. Par. 3e, Reference 1a, defines "...verification inspection to include:

"(1) Army evaluation of contractor's inspection systems to determine compliance with clause 5.e. Standard Form 32 (General Provisions-Supply Contract), or a similar inspection clause contained in the contract.

"(2) Army product inspection performed to measure quality of product offered for acceptance."

4. Par. 4a, Reference 1b, makes contractors "...responsible for controlling product quality and for offering to the Army for acceptance only those items...considered by them to conform to contractual requirements." Clause 5.e., cited in Par. 3e above, makes it a contract requirement that the contractor establish and maintain a system of inspection acceptable to the Government. Par. 3f, Reference 1a, refers to "...records of results..." as integral parts of the contractor's inspection system. Pars. 4b and 5b(1), Reference 1b, place an upper limit on Army verification inspection which "...will not exceed...the total of that inspection set forth in the Quality Assurance Provisions of the specification or contract." Furthermore, the References contain several general and specific allusions to the objective of reducing Army verification inspection when the contractor's quality inspection system is found to be reliable. In particular, Par. 4b, Reference 1b, states: "The extent of Army verification inspection to determine compliance with the Quality Assurance Provisions and other requirements of the contract will be adjusted to reflect the following factors:

"(1) The pertinency, completeness, and reliability of the supplier's inspection records.

"(2) The previous quality hostory of the supplier's product.

"(3) The unit cost of the item."

5. Par. 5b, Reference 1b, states: "When definitive specifications are the basis for procurement, the inspection system of the contractor... will be considered acceptable when quality of produced supplies or services is consistently acceptable, and it includes, as a minimum, the performance of those Quality Assurance Provisions stated in the specification and not reserved for sole performance by the Government." In prescribing the categories of verification inspection, Par. 5d, Reference 1b, states: "The amount of verification inspection will be adjusted to make maximum utilization of the contractor's quality control system and the quality history of the product..."

6. The cited provisions of References 1a and 1b gave rise to the following deductions:

a. Contractors are responsible for controlling product quality and must offer for acceptance by the Army only those items which the contractor considers to conform to contractual requirements.

b. To insure that this responsibility is fully discharged by the contractor, the Government makes it a contractual requirement that the contractors establish and maintain a system of inspection acceptable to the Government; records of inspection results are considered integral parts of the contractor's inspection system.

c. The Government considers the contractor's inspection system acceptable when it includes, as a minimum, performance of all quality assurance provisions of the specification not reserved for sole performance by the Government, and when the quality of material or services produced is consistently acceptable.

d. Army verification inspection, on the other hand, must not exceed the inspection prescribed "as a minimum" for performance by the contractor. In addition, several references are made to the (downward) adjustment of Army verification inspection, depending upon the reliability of the contractor's quality inspection system.

7. From the above, it is clear that contract provisions written in accordance with the references require the contractor to furnish both supplies and a quality inspection service related to the supplies. This inspection service is intended to cover all elements of product inspection which, prior to September 1957, were required of the Government inspector. It is obvious that if the contractor performs this service diligently and with a validity equal to that of a Government inspector then, with assurance that such is and remains the case, the Government inspector could confidently accept contractor inspection results as though he, himself, had performed the inspection.

8. CmlC has always stressed the importance of "feedback" of inspection data as an essential element in controlling quality, and indeed inspection. Product verification insures, among other things,

the existence of a valid body of independent data which, through proper "feedback," can be used as a self-governor to limit government inspection to that required to protect the Government's interests. CmlC considers contractor quality history to be an essential element in quality assurance, useful in preaward surveys and as a check on contract performance. In its implementation of AR 715-20, CmlC uses product verification as an essential, independent estimate of contractor quality history which can also be used to optimize the economics of government inspection consistent with quality requirements. It is noted that this viewpoint dates back to 1944 when CmlC first introduced into its specifications a "quality control" paragraph which stated in effect that if a contractor operates under a system of quality control acceptable to the Government and consistently produces high quality material, the Government might "...modify the whole or in part..." the sampling and testing requirements of the specification. The object, of course, was to reduce Government inspection.

9. The unabridged dictionary defines "verify" to mean "to prove to be true;...to confirm, as by comparison with facts;...to check or test the accuracy or exactness of; to confirm or establish the authenticity of;..." Thus, the objective of the "verification inspection" described in par. 3 above, is to prove, confirm or authenticate (as the dictionary puts it) the validity of the contractor's inspection system by comparing his inspection results with the independent factual findings of the Government inspector. The problem lies in assuring the reliability of the contractor's inspection system. To do this completely required careful check of the administrative and technical phases of the contractor's quality inspection activity, followed by independent product verification inspection by the Government inspector. The validity of the contractor's inspection data is established when no discrepancy is noted in the administrative and technical phases of his inspection work and when no statistically significant difference is found in measuring his inpsection data against those generated by Government verification inspection of the same material.

10. The viewpoint in par. 9 conforms with the definition of verification inspection contained in par. 3e, Reference 1a. It combines Army evaluation of the contractor's inspection system to determine compliance with Clause 5.e. of the Supply Contract, with measurement of quality of the product offered for acceptance. Since product inspection is normally the most laborious single activity of the Government inspector, reduction of Government inspection, with concomitant savings of man hours and dollars, could most appropriately be made here. Therefore, the objective is to reduce product verification inspection to the extent possible through dependence on contractor inspection results provided the validity of these results has previously been assured by thorough-going verification inspection and product verification by the Government.

MEASURING A COMPLEX FIELD OPERATION

K. L. Yudowitch
Operations Research Office, The Johns Hopkins University

We who profess to practice operations research for the Army are confronted with a dual problem: the inaccessibility of the operation and the complexity of the operation. I have chosen to discuss what appears to be a relatively simple military operation: the combat firing of a rifle, or rifle-like weapon. The inaccessibility is inherent in the word "combat" -- there is no available combat operation to provide the proper context. Second, the word "firing" implies the activity of a human being, unfortunately frivulously complex. In order to get meaningful measurement therefore, two procedures are essential. First we must simulate the operation, second we must isolate or randomize the complex parameters. This means that we are obliged to find out how rifles are fired in combat, and then to imitate this repetitively under an appropriate variety of conditions. This will provide I hope a good illustration of the considerations, both quantitative and qualitative affecting the design of such an experiment.

Before proceeding to the design, it is necessary to know just what is being measured. We are interested in the combat effectiveness of rifle fire. The effectiveness component of interest has been designated as the combat accuracy. It has already been determined that combat rifle accuracy is critically dependent upon the error of aim, and negligibly affected by other identifiable errors, generally categorized as interior or exterior ballistics. Our interest in this particular study is further arbitrarily limited to what has been called "snap shooting" -- that is, firing the rifle with a very brief aiming or pointing time.

The study was motivated by a desire to pin down the effects of certain selected rifle characteristics on snap shooting accuracy. These characteristics are four: configuration, recoil, sights, and weight. Much experience and more folk lore gave rise to heated debate on the effects of one or another of these parameters on rifle accuracy. In addition, a fifth effect on accuracy is inherent in the apparent difference between first round and succeeding round accuracy on a single target. Sixth, the skill of the rifleman doubtlessly affects accuracy.

It is clearly a complex matter when one considers that we have undertaken to examine the effects of six possibly interacting parameters. The number of conditions for experiment is inherently quite high. Even if our experiment is extremely rough and examines only two values for each of these six parameters, there are 2^6 or 64 possible sets of conditions. Recall also that we must effect repetitions because of the "random" variations.

5 QUESTIONS

1. Which Conditions Most Difficult?
2. Amount and Confidence of Differences?
3. Which Interactions Negligible?

4. Adequate Parameter Increments?

5. Relative Value of Effects?

How do we now determine a design? Five questions are basic: First, which set of conditions are too difficult or expensive? Second, how much random or statistical variation is permissible? Third, which interactions may be ignored? Fourth, how many values of each parameter are sufficient? Fifth, which effects are of more interest, and which of less interest?

Let's look at the example. For configuration we select the M14 as military standard and (on advice of experts) a popular hunting rifle like the Winchester Model 70. Recoil is M14 standard and reduced (say half) load. Sights for examination are military aperture and hunting type open sights. There is no problem supplying either sight or recoil to both these rifle configurations. Weight is less simple. The Winchester 70 is available in both light and heavy versions -- and the heavy Winchester just about matches the M14 weight. However there is no lightweight version of the M14. This immediately eliminates 1/4 of our 64 sets of conditions, unless we supply special rifles. Our first qualitative question is asked; and in this case special rifles are not deemed economically feasible. Succeeding fire after first round is only of interest for semi-automatic fire. The Winchester is manual only, eliminating another 1/4 of our 64 sets of conditions, unless special rifles are supplied. Again the first question is asked, and special rifles ruled out. Our design has already been reduced to half, 32 sets of conditions.

SETS OF CONDITIONS

4 Sets			8 Sets		
Config.	Weight	Mode	Recoil	Sights	Skills
(M14	Std.	Sing.)	High	Open	Exp.
M14	Std.	Mult.	Low	Aper.	Mark.
W70	Std.	Sing.			
W70	Light	Sing.			

The eight combinations of recoil, sights and skill values apply to all remaining sets of configuration, weight and mode. As the M14 single mode data are included in the multiple mode data, the M14 single mode set may be deleted without loss, leaving only 3 times 8, or 24 sets of conditions. It is only necessary to identify the first bullets fired at each target -- easily done by painting those bullets. The paired comparisons provided by these imperfectly balanced 24 sets of conditions are listed:

COMPARISONS

CONFIGURATION (M14 vs. W70)	8 (Recoil x Sights x Skill)
WEIGHT (Std. vs. Light)	8 (Recoil x Sights x Skill)
RECOIL (High vs. Low)	16 (Config. x Wt. x Sights x Skill)
SIGHTS (Open vs. Aper.)	16 (Config. x Wt. x Recoil x Skill)

We see that without repetitions, these conditions provide 48 comparisons: 8 configuration, 8 weight, 16 recoil and 16 sight. Having answered the first of the 4 questions, and thrown out half of the possible conditions as impractical, we next ask about statistical reliability.

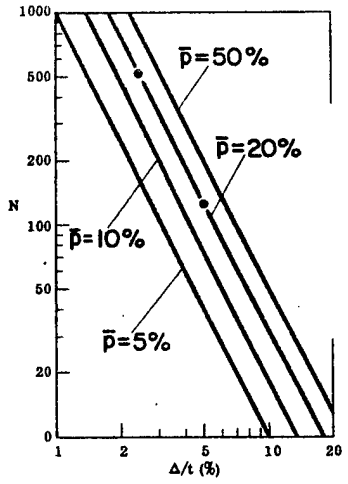
How many repetitions are needed to provide sufficient confidence in our results? This is clearly dependent on how large a difference we are seeking to identify. If the rifle accuracy is characterized by a hit probability p_x, and the changed condition results in p_y, we may note the change in p as Δ. It is then a simple exercise to relate this change to the number of measurements N and the familiar student's t, measuring confidence level (See graph.).

Estimating an average rifle hit probability of 20%, and an objective of identifying a 5% change in this probability with 95% confidence ($t = 2$), we get a requirement for 500 measures. If we are satisfied with identifying a 10% change in hit probability, only 125 measures are needed.

If we use 7 firers of each skill level, and each man fires at 7 targets, we should need to run through only 2 or 3 times for a 10% difference $[125/(7 \times 7) \simeq 2\frac{1}{2}]$. A 5% difference required 10 replications $[500/(7 \times 7) \simeq 10]$.

The decision on number of replications calls for consideration of our third question. The third question is the educated guess about which interactions are negligible. Surely the 8 comparisons of the configuration are not all independently different. Similarly for the 8 comparisons of weight and the 16 comparisons each of sights and recoil. If there were no interactions at all, we could expect to have adequate statistical reliability to identify a 5% difference from a single run-through. Clearly we want somewhere between 1 and 10 replications. Value judgments finally determine the number of replications.

Experience predicts experimental running time. If it is desired to complete the experiment in one week in the field, one comes up with 3 to 4 replications. Of course the multiple round modes of fire supply more



V7947

$$N \simeq 2\bar{p}\bar{q}\Big/\left(\frac{\Delta}{t}\right)^2$$

$$\begin{cases} \bar{p} \equiv (p_x + p_y)/2 \\ \bar{q} \equiv (1 - \bar{p}) \\ \Delta \equiv (p_x - p_y) \end{cases}$$

data per run (by a factor of the number of rounds fired per target). Thus a reasonable design permits 4 single mode replications (64 runs), and perhaps 2 multiple mode replications (16 runs).

As value judgment indicates this experiment is worth about 1 week, and statistical reliability obtained in that time is just about adequate, the fourth question regarding the number of values per parameter is answered; No. Any refinement of the effect of more than 2 increments of the four basic input parameters is best postponed to a succeeding effort.

Our fifth question evaluates the several comparisons which the experiment makes. If there is especially great interest in the effect of recoil for example on the current military standard rifle, we might incorporate extra runs for both recoils with Marksmen firing the M14 with aperture sight in multiple mode. Similarly, reduced interest in one of the parameters might dictate deletion of some conditions. However deletions must be made with care, for each condition is used in an average of two comparisons, and this advantage of the semblance of balance that remains in our design is quickly lost by deletions.

A further incidental advantage of the minimal experimental design is its relatively lower susceptibility to biases of learning, fatigue, weather changes, etc.

My purpose has been to illustrate how practical considerations of time, value and cost impose over statistical considerations to define an experimental design. The result is generally unbalanced in a statistical sense, but balanced in an operational and value sense.

THE CONDUCT OF MILITARY FIELD RESEARCH ON A SHOESTRING

Andrew J. Eckles, III
Operations Research Office, The Johns Hopkins University

There is, I believe, a certain poetic justice in the fact that this paper is the last one to be given at this conference. For after all, our subject matter deals not with what we would really like to have, if we "had our druthers;" but rather with a last resort, stop-gap method of doing our necessary field research.

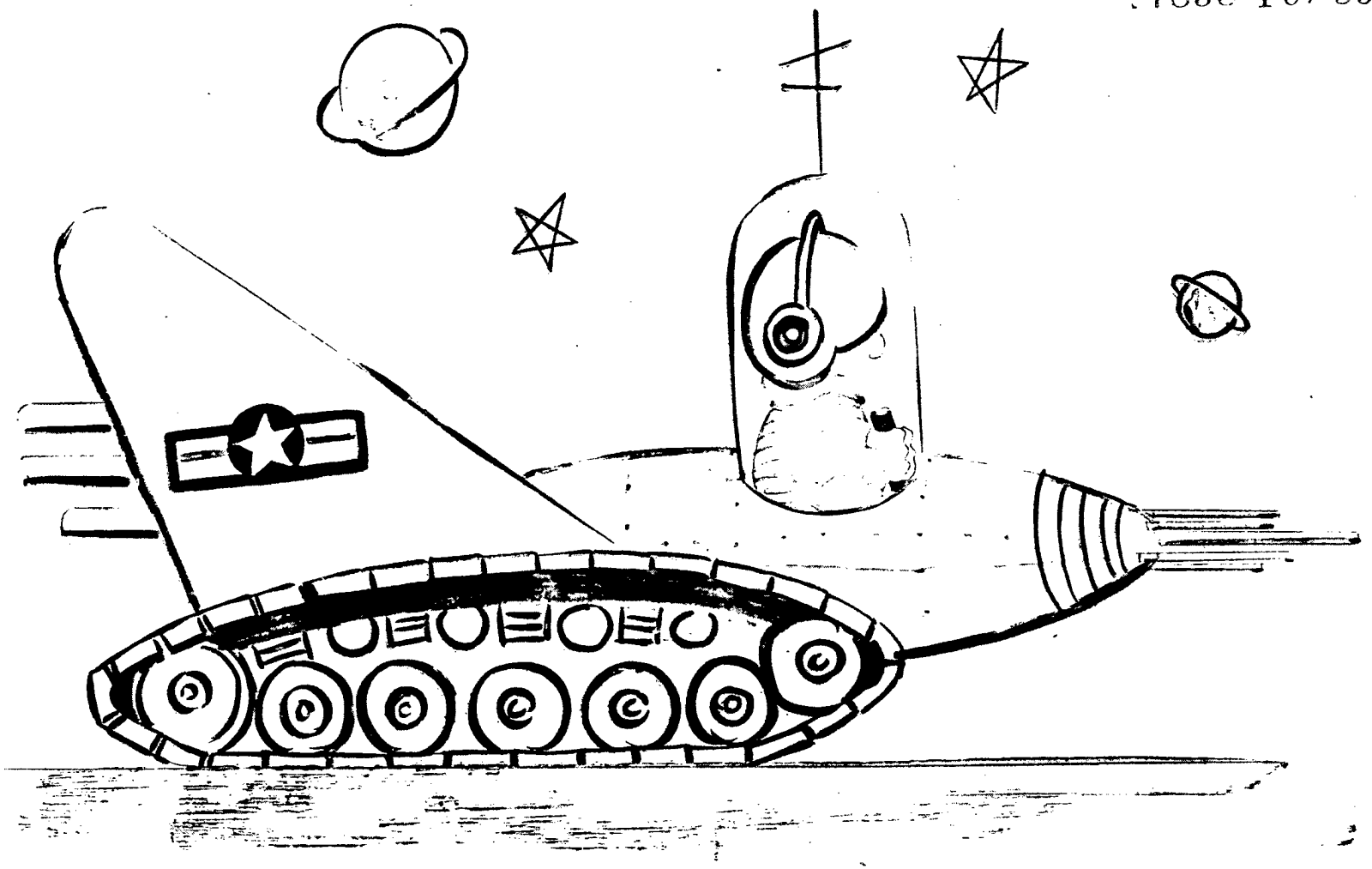
In recent years, our methods of wargaming and other analytical techniques have increased in accuracy and reliability to such an extent that their need for reliable input data have by far outstripped all available facilities. To some extent, this urgent need for valid performance data under actual field conditions has been recognized; and as a result, CDEC, an agency admirably suited for the conduct of sophisticated, controlled field experiments in the area of tactical operations was established. But, as most of you know, the conduct of operational field experiments is not only time consuming, but extremely expensive in manpower and equipment as well as money. Therefore, all of our field research facilities together, with the means available to them, cannot hope to keep pace with the increasing appetites of our model makers for the basic data upon which to perform their manipulations.

In most cases, our military analysts have been forced to resort to guesstimate, or even worse, manufacturer's specifications, for the vast majority of their basic input parameters--which, of course, regardless of how sound the model, reduces most results to little more than science fiction.

As a result of this shortage of reliable input data obtained under adequate field conditions, we at ORO have been investigating some short-cut methods of conducting small scale operational field studies. In this paper, then, we will present one possible solution to the cost factors involved in the conduct of complex field experiments. We will discuss, literally, the conduct of such studies on a shoestring.

Perhaps I should emphasize that what we are discussing here is nothing new--in fact, the basic ideas have even been presented at one of these conferences a few years back. But we have made some efforts to codify our techniques for greater efficiency, and have even coined an expression "SYMBION," to connote the underlying principles. Essentially, the concept of Symbion is to superimpose experimental designs and data collection techniques upon carefully selected, and at times modified, phases of the Army training Programs. This provides us with a means whereby we can tap the vast reservoirs of manpower, equipment and ammunition expended in the normal traning cycle as a potential source of valuable weapons systems performance data.

We have recently completed several studies utilizing this idea of Symbion, and presently have another under way. So before discussing the principle itself, with its good and bad points, I would like to describe briefly two of these studies.



Good - if possible

Slide 1

Bear in mind, now, that this work which I shall describe is not exactly what we would like to have done. But we found ourselves in the condition of the poor soldier with mighty ambitions. For example, his ideal weapons system might be something like this (See Slide 1), complete with air conditioning and fully stocked bar. But there can be certain economic and other obstacles in his way. And yet, he has a job to do--with whatever equipment he has available or can scrounge. So, like this hypothetical soldier, in order to get some place to do something, we had to be satisfied with a little more primitive situation. (Slide 2)

The first study which we will discuss was designed to investigate the effectiveness of tank high explosive fire against hastily prepared antitank gun positions. The data collection phase of this study lasted from September 1958 through March 1959; though the actual time involved in data collection was less than three weeks. To some extent this was an unusual situation in that we wanted to test our concepts of Symbion as well as obtain the data we sought. To do this, we threw the entire burden for the conduct of the study on the Fort Stewart, Ga., compliment. Our only participation consisted of advice and criticism, and in the development of the initial equipment to be used, which was then duplicated and used by post personnel.

Our first step, in this case, was the development of a realistic antitank gun target which would respond appropriately to H. E. ammunition, and yet which would be realistic enough, cheap enough, and easy enough to use that the Army Post itself would provide all of the equipment which was needed for the conduct of the experiment. Part of this problem was easily solved by modifying a previously used gunfire simulator so that it could be built by Post Signal personnel utilizing only surplus drone target parts which had been salvaged. Our Electronics Laboratory built the first such system, and the remaining ones were then built by Post Signal. To obtain a realistically killable target to serve as gun crewmen, we developed one which consisted simply of a toy balloon inserted an a canvas bag. Such a target possesses amazingly realistic characteristics. After our first run, Ft. Stewart constructed their own target bags, but until this killable target was adopted by the Army as a training aid, our Office still had to purchase the balloons-which, oddly enough, were not standard army equipment at that time and difficult to justify to the comptroller.

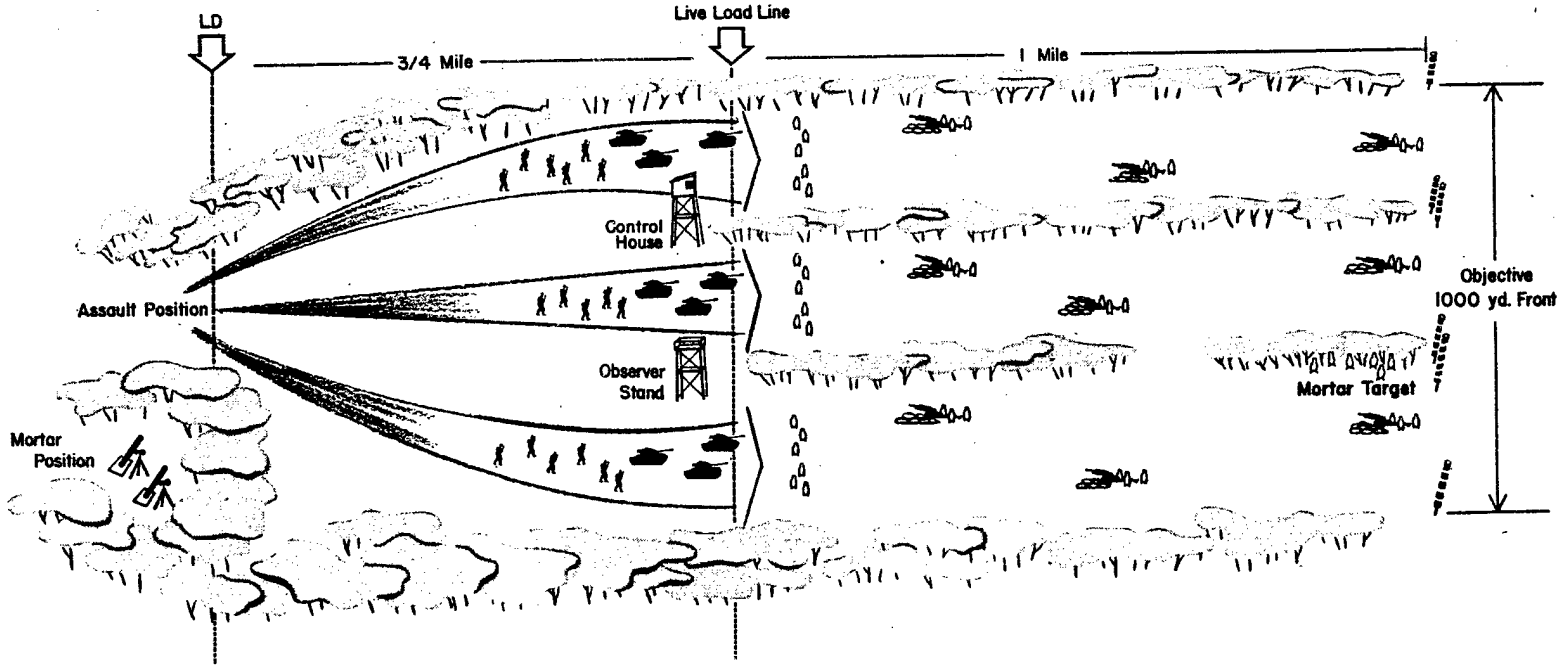
After we had developed the prototype target system, an experimental course was set up by the Post, following our proposed plans very closely. The field setup looked something like this. (Chart of experimental setup--Slide 3.) Essentially, this was a Company live fire problem. However, for convenience of data collection, the problems were conducted as three simultaneous platoon courses. As shown here, there were three separate lanes, one for each platoon. Each lane consisted primarily of three antitank gun targets located in the following range brackets: 300-400 yards, 450-550 yards, and 1200-1500 yards.

Each antitank gun target consisted of the following items: a mock antitank gun; a gun fire simulator which would detonate, on command, up to three charges of TNT (representing the AT gun firing); and a crew consisting of our newly developed "killable" balloon targets.

Often
we must be more primitive

V785210/59

Slide 2


LD
Live Load Line
3/4 Mile
1 Mile
Control House
Assault Position
Observer Stand
Mortar Position
Mortar Target
Objective
1000 yd. Front

Slide 3

TABULATED RESULTS

V7951

Crew Number	Gun positions		Crew members		Rds Fired	Kill probabilities per round	
	Available	Killed	Available	Killed		Gun position	Crew member
3	9	4	27	5	54	.07	.09
4	24	6	96	20	144	.04	.14
5	12	5	60	5	72	.07	.07
6	12	9	72	20	120	.07	.17
Totals	57	24	255	50	390	.06	.13

(Unfortunately, we did have one serious fly in the ointment. This was the availablity of balloons. We started with four small targets and two large targets making up each AT gun crew. However, the number of crewmen serving each gun varied from phase to phase of the problem as we ran out of balloons, and as our supply of funds varied for the purchase of additional ones. Indeed this study was conducted on a shoestring.)

In addition, to enhance the realism of the test situation, overhead mortar fire was delivered on the objective, and infantry supported the tanks in the attack. Additional killable targets were utilized for mortar targets and for infantry targets. These gave us, purely as a by-product, some information on the effectiveness of mortar and .30 caliber firing.

We have, to date, obtained fairly reliable information on twenty-one different tank platoons that have gone through this test course. Our primary information, concerned with the kill probability of a round of H.E. fired against an unarmored gun crew, resulted from the expenditure of a total of 390 rounds of 90-mm H.E. ammunition. The total cost of this information to our Office, not counting the time of one analyst, and development of prototype equipment, amounted to somewhere between $1.50 and $2.00 per data point.

Had we conducted this very same study as a standard field experiment, again not counting the time of one analyst and development of prototype equipment, the cost per data point would have been somewhere around $300.00.

A rough summary of the primary results in this study are shown on this chart. (Slide 4) Since this is more a report of methodology than results of a specific experiment, I won't spend too much time giving you our results, but briefly this table is interpreted as follows:

The first column gives the number of killable targets in each gun position, for a given phase. The second column gives the number of gun positions available in each phase (always in multiples of three, since there were three gun positions per lane). The third column gives the number of gun positions killed-that is, a gun position was considered killed if one or more of the crew members serving that gun was killed by an HE fragment. The fourth column gives the number of crew members available to be killed in each phase (found by multiplying the number of crew members per gun by the number of gun positions). The fifth column gives the number of crew members killed by HE fire. The sixth column gives the total number of rounds fired during a given phase. The seventh column gives the probability of a given round inflicting a kill on a gun position. And finally, the last column gives the probability of a given round inflicting a kill on a given crew member.

A quick glance at the kill probabilities would indicate that an AT gun position has a very good chance against an individual tank firing one round of HE. However, as indicated in the first two columns, when an objective of this type is attacked by three tanks, a given gun position has only a little better than a 50% chance of survival from the tanks 90-mm HE fire. The use of the tanks sub-caliber weapons would reduce these chances of survival even further. Also, observation of the actual

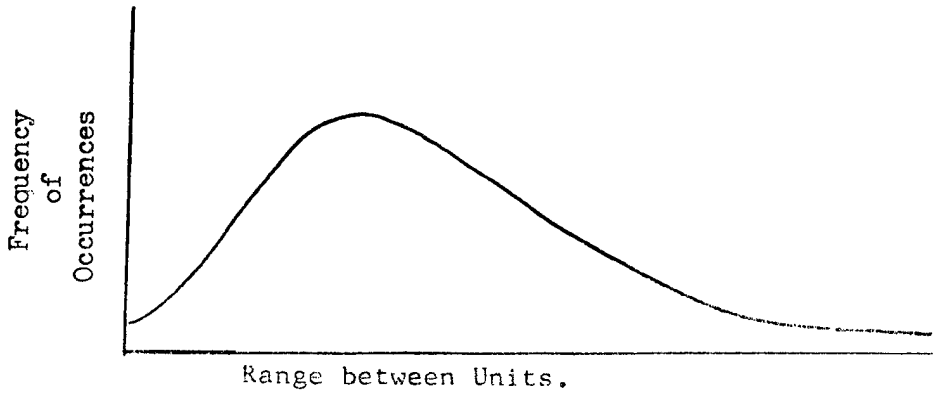
situations would lead one to think that rigid fire discipline would decrease the survival rate of AT guns even more.

I would now like to discuss another project which demonstrates how these shoestring techniques can be utilized to collect usable basic data. This study is presently being conducted by the Communications Group at ORO, and the data collection phase will continue, at periodic intervals, through next Spring.

It is well known that adequate communications have always been a problem in the wars of history, and will no doubt continue to be even more so in wars of the future. In an effort to cope with these problems, it has been usual for the soldier to ask constantly for radios with longer range, greater reliability, etc., but communications threaten to become the tail that wags the dog. However, there is a group at ORO presently working on the anticipated requirements for communications in the next time frame--and perhaps for the first time a realistic attempt is being made to try to determine just what communications facilities will be needed, and what can be discarded as dead weight or excess clutter on the air-waves.

Considerable information concerning the future need in this area has been obtained by means of wargaming. However, there is as yet no data available which could show a correspondence between the results obtained from our analytical models and the results which can be expected in the real world. The present study, then, is an effort to provide data which would help define the relationships being studied.

I have time to show only one facet of this study, as an example of its objectives. One measure obtained from analytical techniques which can help provide guidelines for communications requirements is a frequency distribution showing range between units which need to communicate. Such a hypothetical distribution might look like this:



But one very serious question is, just how valid is such a distribution which is obtained only from mathematical models; is the shape even similar to what might be found in a real life situation? Is the total message count realistic?

It is our hope that this "ongoing" study will provide data which could help answer these questions.

Our subjects, in this case, will be individual tank battalions with supporting infantry. They will, to the greatest extent presently possible, be placed in a quasi-combat situation for 72 hours, in which they will engage an aggressor force under a variety of situations. During this period of test, the ORO group will monitor the battalion net and record each radio transmission together with the distance between the units concerned. From this information, a frequency distribution will be made as from the war games. The correspondence between these curves should provide at least one base point for estimating the correspondence between the results of our war games and what might conceivably happen in the real world. Of course much additional information will be obtained--but the above is a prime example. By this spring, we hope to have data from at least four battalions.

At the present time we do not know just how many data points will be obtained from each battalion. However, if we assume that we obtain only about 500 data points from each battalion, by using the techniques of SYMBION, the approximate cost from research funds will be somewhere in the order of $2.00 per data point (again not counting the analysts' salaries, which would be constant with either method). On the other hand, to conduct an identical situation as a standard field experiment (and, by the way, the situation which we would have set up is not very different from the one now being conducted as a training program) the cost of this information would be somewhere in the neighborhood of $500.00 per data point.

And besides this, we could never hope to obtain the number of different subject battalions which are easily available to us in the normal training cycle by merely extending our data collection phase.

Now that you have seen some of the capabilities of this "cheap-skate" method of data collection, I would like to discuss some of the basic principles behind the technique.

Essentially, the principle objective of the concept of Symbion is to increase the efficiency of our available research facilities--that is, to get the very maximum out of every research man-hour and research dollar (and it is very difficult to say which is more scarce at present).

The real value of this technique is shown whenever we attempt to obtain performance data under quasi-combat conditions as opposed to manufacturer's performance data or proving ground type data. For it is here that the cost of operational field studies mounts to prohibitively high figures--and yet it is also here where the soldier is most willing to work closely with the scientist in order to increase as much as possible the realism of his training programs.

The procedures for conducting a Symbion type project are relatively simple. It involves first the selection of an appropriate phase of the Army Training Program which will provide the basic situations required for

our experiments. Every effort is then made to develop equipment which will provide the necessary realism to qualify those situations as quasi-combat. If the equipment developed is "practical" enough, from the over-all military viewpoint, then the cost of this equipment can even be borne by the training program, and not out of research funds; however, in some cases it will be necessary to utilize research funds to furnish this equipment. All of the data collection equipment will normally have to be provided by and operated by the research agency.

Let us now examine some of the differences between a specifically conducted experiment and one conducted by means of these "shoestring" techniques. The primary differences can (but not necessarily must) exist in the following areas: experimental control; precision of measurement; time to obtain the required information; sample sizes and representation; and problems involving the general area of experimental design. I will touch briefly on each of these five areas.

Experimental control is a major problem in any complex operational field problem--and an area in which many concessions must be made due to such necessary factors as safety, limited terrain, etc. But these problems must be faced no matter how we conduct our research. It is true, however, that there will be times when, under a Symbion type program, we will not have the control which we would like to have over all of the factors which may affect the results. Often, once we have initiated the desired quasi-combat courses and installed our data collection equipment, we are then reduced to the position of the astronomer--that is, we may be limited in the actual manipulations which we are allowed to perform, but must obtain most of our data by simply observing and recording events as they occur. Such, for example, is the case in the communications study which was described earlier. (But there, to a large extent, we would not exert any more control even in a specifically set up experiment--largely due to the fact that as yet we do not know exactly what factors we wanted to control in this particular type case.)

This, at times limited control, then leads us to question the precision of the data obtained by these methods.

As for precision, data collected by means of a Symbion-type program can run the gamut of quality just as those obtained by any other means. In the great majority of cases, if the basic program is properly organized and set up, the resulting data will be every bit as accurate as that obtained by any other means. But, again, there will be times when, using our quick and dirty techniques, we will have to accept data which are not as precise as those which could be obtained from a specific experiment. It then behooves us to weigh carefully the additional cost (in manpower, equipment, and time as well as dollars) of each increase in precision required.

When evaluating this additional cost, it should be remembered that, unlike most scientific research, where the results of our efforts often form lasting foundations for pyramiding results, military research to a very great extent is dated and all too quickly becomes obsolete as weapons systems and the characteristics of war change. Thus much of what the

military researcher does has only temporary value. Knowing this, we must consider the cost of our work in the light of its often time-limited usefulness. It is therefore usually much wiser to be satisfied with "broad-brush" type answers in order to have timely solutions to our problems, and then spend our remaining efforts on other situations that are also crying for solution. We just cannot afford to waste precious time in obtaining data that are somewhat more accurate, but are useful for only a short period. We must of necessity develop short-cut methods of obtaining as rapidly as possible the data we require.

We have now somewhat sneakily entered the area of time required to obtain our data. And it is here that our "quick and dirty" techniques can prove to be very quick indeed. A developed Symbion-type program is perhaps the only hope we have in the foreseeable future of providing the vast quantities of performance data required in any realistically acceptable time frame. A completed series of "quasi-combat" training programs of the type which we have described would provide the military scientist with full-blown, completely equipped "laboratories in the field" which would be readily available for his needs as they develop. These "laboratories" would be staffed and run by competent and experienced soldiers already acquainted with the equipment required for a quasi-combat course, since it would have become part of their everyday performances. In addition, they would have become even somewhat inured to the often rigid demands of the scientist. With such a program, the scientist would be prepared to examine rather complex problems with a minimum of delay, and in most cases obtain some of his required data within months rather than years of an expressed need.

We now come to the area of sample size and sample representativeness. In most large scale field experiments, we are forced to limit severely our sample sizes, and make repeated use of "test-wise" troops. This can, to a very large degree, limit the extrapolations which could be made from the results of our studies to "troops in general." In many cases, then, it would seem to be very advantageous for us to utilize the Army Training Program in order that our samples be large, and that they more closely resemble the army-wide population in such important factors as degree of training, motivation, intelligence, etc. For example, in a specific experiment, a sample of four or five test-wise experimental companies might be considered a large sample; but in a Symbion type program, we could consider a sample of ten naive companies relatively small.

(As an aside, it should be mentioned that we can to some degree compensate for lack of control in this type study through increases in sample size.)

The final problem area of major differences is that of experimental design. The design problem is often, it is true, made much more difficult in our shoestring type situation than in specifically conducted experiments. More often than not we will have to be satisfied with the more crude, less sensitive designs and their companion analytical techniques. We simply have to utilize designs which can be fit into an already complex and crowded training schedule. But when the final results are compared, we will find in most cases that we have lost nothing of practical significance (as

Slide 5

Slide 6

distinguished from statistical significance). And this is especially true when we are faced with the choice of either no data at all (other than guesstimates) or data obtained under less than ideal conditions.

The completed program of this type will always be conducted by the training officer, and the scientist is reduced to the role of advisor and observer. But it is surprising how much common ground can be found between the soldier and the scientist especially when, after a long day in the field, they stand side-by-side at the bar. We have found that the participating soldiers are not only willing but eager to assist us at every opportunity--and almost every suggestion made to implement scientific validity has been followed as far as safety requirements and available resources have allowed.

And I should add that we have also found that our most vociferous, and often most helpful critic has been some frustrated soldier trying to persuade a piece of balky data-collection equipment to operate in the rain, and do his own job at the same time. But the soldier, too, is interested in objective measures of his performance under combat-type conditions, and will consider his extra efforts well worth while.

In summary, then, we feel that these economic methods of data collection do enable us to obtain usable weapons systems performance data under quasi-combat conditions with a minimum expenditure of research time and money--both of which are extremely limited. In addition, the training officer is quite eager to participate in any program which he feels will add realism and motivation to his programs.

A judicious utilization of appropriate phases of the Army Training Program can, then, enable us to conduct much of our necessary field research on a shoestring. And, in addition, the participating soldier gains as much of immediate value as does the scientist when they join forces.

Before closing, I would like to add one word of warning. We cannot overemphasize the fact that the research methods which we have described here are not used merely by preference. They are "make-shift," "stop-gap" techniques--which, nevertheless, provide us with a source of usable data which would otherwise be unobtainable. Perhaps I could demonstrate this best by means of a cartoon. (Slide 5)

Like this gentleman, we often have our goals clearly in mind, and are well aware of the more sophisticated and refined techniques which we would like to use. But these techniques do cost money, and they do take time--and sometimes we find ourselves short of both. In such cases, then, we might find that the more direct approach used by our ancestors, even though somewhat crude, could work wonders. (Slide 6).

PROPOSAL FOR FIELD CALIBRATION OF A TRACKING RADAR

Victor B. Kovac
RCA Missile Test Project, Patrick Air Force Base

PURPOSE.

The purpose of this experiment is twofold: to determine the feasibility and adequacy of automatic tracking of sun data to evaluate several types of angular misalignments in the radar mechanical system; and to investigate the feasibility of obtaining ship's position accurate enough to serve as origin for data reduction purposes.

INTRODUCTION.

The coordinates of a missile's position recorded during radar track at 0.100 second intervals are azimuth, elevation, and range. These coordinates and time are channeled directly into a digital computer for "real time" data presentation, and to a tape recorder for subsequent data reduction. The angular resolution of the AN/FPS-16 Radar is 0.05 mil and the tracking precision is 0.1 mil. Absolute errors are not large, but are believed to exceed 0.1 mil. Hence, it is desired to remove systematic errors so that ultimate, smoothed data is not only precise but accurate.

Owing to the use of several radar stations to track long range missiles, it is essential that they are accurately aligned with the common geodetic reference, so that overlapping spans in time also agree in position. During the data reduction process, small misalignments can be corrected if they are known. Evaluation of these misalignments in the mechanical system as a function of the radar's azimuth and elevation orientation is one approach to the solution.

During tracking of a missile, there must exist small deviations of the target which the radar attempts to correct automatically, or else it could not track. This dynamic characteristic of tracking tends to produce angular deviations in data which are difficult to separate from the small alignment biases. Furthermore, refraction (bending) of the radar beam in the atmosphere is a complex function that has to be accounted for. Partial correction of refraction errors in data has been effected, but the oscillatory nature of the dynamic error requires further study.

Figure 1 (at the end of this article) is an example of seven seconds of typical raw elevation data during an early portion of a missile track. It shows:

1. Time (abscissa) at 0.100 second intervals
2. Raw elevation points (radar)
3. Refraction error (from separate sources)
4. Reference standard (position data from cine-theodolite cameras near the target translated into elevation at the distant radar origin.
5. Elevation error (bias between a radar data point corrected for refraction, and the theodolite point).

The elevation scale has been purposely exaggerated in order to illustrate: The magnitude of refraction error, the oscillatory nature of the dynamic track, and the difficulty of evaluating angular biases in which we are interested.

Figure 2 illustrates the three orthogonal axes of the radar's mechanical system. A deviation between the encoder bearing and the true bearing results in a constant azimuth index error. A small rotation about the OX or OY axis (or both) corresponds to turntable mislevel or tilt. An elevation axis tilt with respect to the horizontal is called a standards error.

Electrical (or optical) axis deviation from normal to the elevation axis, constitutes a departure from a truly orthogonal system. In this instance, the deviation of the electrical axis from the plane of the elevation circle causes an error in azimuth. Dial eccentricity denotes eccentric deviation of the dial shaft from the true center.

MISALIGNMENTS and THEIR EFFECTS (Figure 3)

Mislevel.--Misalignment of the radar's vertical axis with respect to the local vertical results in a turntable mislevel from the horizontal. The resulting elevation error is sinusoidal, having a period equal to one complete rotation in azimuth, an amplitude equal to the angle between turntable and horizontal planes, and a phase (θ) defined by the direction of the intersecting planes and North. (Equation 1). Two components of error denoted by α and B at 0° and 90° azimuth, respectively, also describe a particular sine wave (Equation 2). The effect of a tilted turntable on on azimuth error is a tangent function of elevation as shown in Equation 3. Thus a 1-mil tilt results in a sinusoidal azimuth error with 1-mil amplitude at an elevation of 45°. In Figure 4, the increasing magnitude of the amplitude is indicated by the dash-dot line.

Standards Error.--This error is caused by a misalignment of the elevation axis with respect to the horizontal, and is analogous to mislevel, except that it is independent of azimuth. This is so because the standards, which support the elevation axis, rotate with the turntable and thus always present the same aspect with respect to the target. Both standards error and mislevel may be evaluated by direct measurement, using the boresight telescope and leveling bubbles in each case. Both direct methods have a resolution of 0.05 mil. Evaluation of component errors by independent means simplifies the resolution of the total error, (i.e., reduces the number of unknowns).

Electrical (Optical) Axis Deviation.--A line which deviates from the elevation circle by an acute angle (ℓ) generates a flat cone during a complete revolution about the elevation axis. Thus, the radar's electrical axis may deviate from the elevation circle. The net effect is an error in azimuth. An axis deviation of 1 mil results in an azimuth error that is a secant function of elevation, as shown by the dashed line in Figure 4.

Bending Error (Antenna Unbalance).--The bending effect due to antenna unbalance is nearly compensated by a pair of spring equilibrators mounted on the radar standards. The cosine function of elevation listed in Equation 5 (Fig. 3) is intended to match the true error that may arise at various elevations. However, the true condition is unknown, hence the form of the mathematical expression as well as the unknown coefficients of bending error require careful analysis.

Dial Eccentricity.--There exists a gear train between the respective azimuth and elevation shafts and the data encoders, so that the successive contributions of error due to eccentricity in each gear result in a complex wave. According to Barton, (1) the measured values in azimuth and elevation are 0.03 mil rms and 0.005 mil rms respectively. Equations 9 and 10 are not applicable directly because they are correct only for cases of one dial. (2). Their form may be of use in the analysis of the residuals.

Index Errors.--Index errors are constants. They denote index deviations from North and from the horizontal.

METHOD

Optical and electronic instruments are subject to error with regard to absolute position of a target in space. Unless the geometry of the other instrument and target is favorable, so that the bias as seen from the radar is negligible, then such instrument cannot be used as a standard. Thus, for a long-range radar, there are few instruments or conditions which offer sound comparison. (Actually, an effort is now being made to track aircraft simultaneously with radar and still cameras which use star background as reference points). Radar, being a point source, requires that calibration be performed over a hemisphere centered at the radar. Furthermore, the coordinates should be in absolute agreement with the geodetic system. Finally, in order to recuce angular error to a minimum, the distance to the reference points should be very long, and the elevation angles should be high enough to avoid multipath effects ($>5^{\circ}$).

In view of the considerations above, it appears that direct measurement of celestial bodies may be the best means to calibrate radars. Star calibration using telescope (or camera) mounted on the elevation axis has proven successful (3). Recently, automatic tracking of electro-magnetic radiation from the sun has been performed, so that direct comparison of computed azimuth and elevation with raw radar data is feasible.

Investigation of the range of azimuth and elevation coverage available from sun track indicates that this source should furnish most, if not all information needed. Some auxiliary points, such as a surveyed reference (Boresight Tower), and the star, Polaris, should be included to complete azimuth coverage and to check electrical-optical axis collimation. Observations of two celestial points at azimuths 60° apart may be reduced to geodetic coordinates sufficiently accurate to serve as a fix for the radar origin in data reduction. Thus, our purpose is to determine whether a number of bursts of sun track, combined with a few auxiliary points, furnish sufficient information to provide angular calibration of the radar.

To test this, the proposed experiment requires an additional, (optical) star calibration to serve as standard. In any case, it is essential to find a dependable means to determine the magnitude of misalignments present in a radar.

TEST DESIGN, REDUCTION, AND ANALYSIS.

A discussion of the reference frame, computation of coefficients and evaluation of the residual errors is in order. First, the range of azimuth is 360°, or a normal sector of 180° plus an equal "extension." Then, the elevation ranges from 0° to 90° , and the "dumped" position from 90° to 180° (Figure 4). Thus we can obtain two observations, or two coverages of the same hemisphere. The sun track will furnish points in the two southern azimuth quadrants with peak elevations ranging from about 60° to 80°, depending on the station's latitude and the season. Low elevations, say below 20°, are undesirable, being subject to considerable refraction error.

Reference to the equations of component errors indicates that they are well-behaved functions of radar orientation. Bending error, being one that cannot be readily isolated, requires that the instrument be dumped. As a consequence, it is necessary that sun tracks be performed in the normal and dumped positions. In order to provide valid standard, it is evident that at least one optical calibration pass should be made completely overhead, (e.g., North-South), and that two azimuth passes at constant elevation (say, E1 = 45°, E2 = 71.6°) should be made.

The tentative scheme for evaluating the coefficient and making an analysis of the residuals is as follows: The array of azimuth and elevation errors will be treated as two separate problems. The elevation errors will be treated first in order to establish the mislevel coefficients. These can be checked against actual measured values using level bubbles. There should be sufficient points to fill four 1 x 4 matrices (m = 16). This means that there should be at least four sun observations before noon (repeated with dumped observations) and another set after noon. Each observation consists of three 10-second bursts of data spaced one minute apart. The reason for this is to permit computing three Solar positions at one minute intervals so that their slope may be known. Also raw data trend and bias can be evaluated more readily from three bursts than from one, particularly since the fitting of a trend is needed to narrow down the radar's wander about the Sun's center.

The azimuth errors will be entered in matrix form and one known coefficient removed (either b from measured standards error, or one of the mislevel coefficients determined by the elevation data). This will leave a 4 x 4 matrix to be solved. Actually, there will be four such sets.

Now a check on agreements among the four separate sets of solutions will be made. First, a least-squares solution will be attempted. The entire array of data will then be treated to removal of the systematic errors. The residuals will then be inspected to determine whether cyclic effects are still evident. Here we can effect a review of bending error

and possibly eccentricity. The remaining residuals, in factorial form, should reveal whether further adjustments in the index errors need be made. Analysis of variance should yield a clue to trends as a function of azimuth and elevation.

As an alternative measure, the factorial format of the four solutions (four quadrants) should reveal complimentary reversals. These can be interpreted as adjustments to be made in the coefficients. Once the coefficients show reasonable agreement, a final review of the residuals can be made with regard to oscillations due to bending effects and eccentricity.

The final comparison of coefficient magnitudes and experimental error is between sun track and optical star calibration results. Here the effect of collimation error between electrical and optical axes may have to be taken into account.

Selecting sun track data at approximately 150° and 210° azimuth, a solution of the radar's geodetic position will be obtained for comparison with surveyed values. If the agreement is good, this means of position indicating may be of use for ship-borne radar stations. This procedure may furnish a check on the magnitude of refraction errors.

CONCLUSIONS.

1. Samples of automatic sun track taken during a day should furnish sufficient information to permit evaluation of the angular errors present in the radar's mechanical system.

2. Star calibration using the boresight camera can be used as standard for determining the accuracy and adequacy of the automatic sun track data to evaluate misalignments.

3. An error surface developed in terms of the radar's azimuth and elevation orientations can be reproduced by analytical expressions suitable for correcting data by machine process.

4. As a possible by-product, two day-time sun orientations may be used to establish a ship's geodetic position.

REFERENCES

1. D. E. Barton, "Instrumentation Radar AN/FPS-16 (XN-1), Evaluation and Analysis of Radar Performance," R.C.A. Missile and Surface Radar Dept, Moorestown, N. J. Jan., 1959.

2. H. Schmid, BRL Report No. 764, "Systematic Errors of Cine-Theodolites," Aug., 1951.

3. K. E. Pearson, "Evaluation of the AN/FPS-16 (System Nr 1) At White Sands Missile Range," Tech Memo 606, U. S. Army Signal Missile Support Agency, White Sands Missile Range, New Mexico, Feb 1959.

Fig. 1. RAW RADAR ELEVATION DATA VERSUS TIME

Position Data from Close Cine-Theodolites (Smoothed), Translated into Elevation at Distant Radar Origin. Magnitudes of Refraction Error.

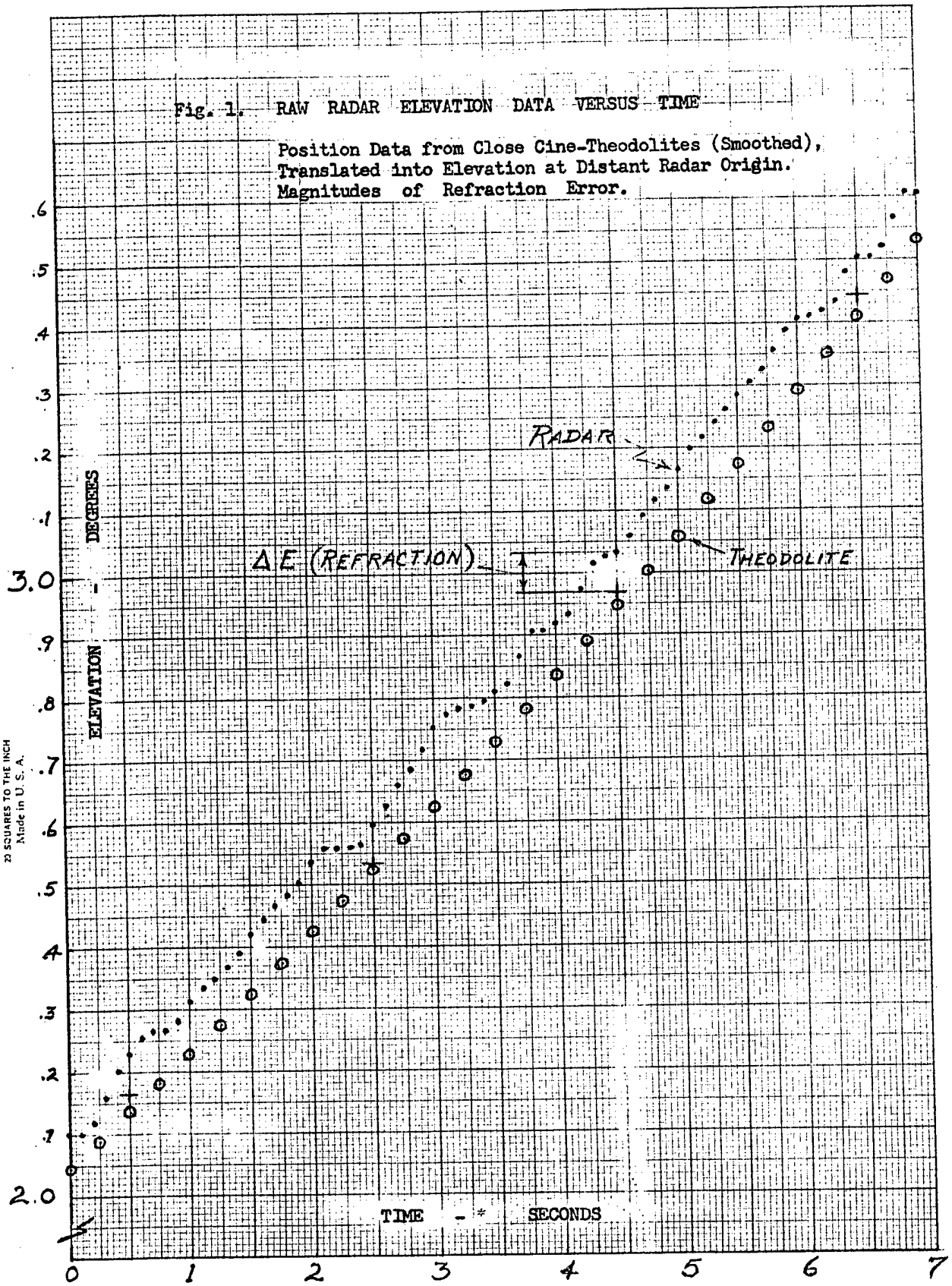


Fig. 2. SCHEMATIC DIAGRAM of a RADAR'S SYSTEM of AXES

Showing Correct Alignment With Geodetic North And Local Vertical.

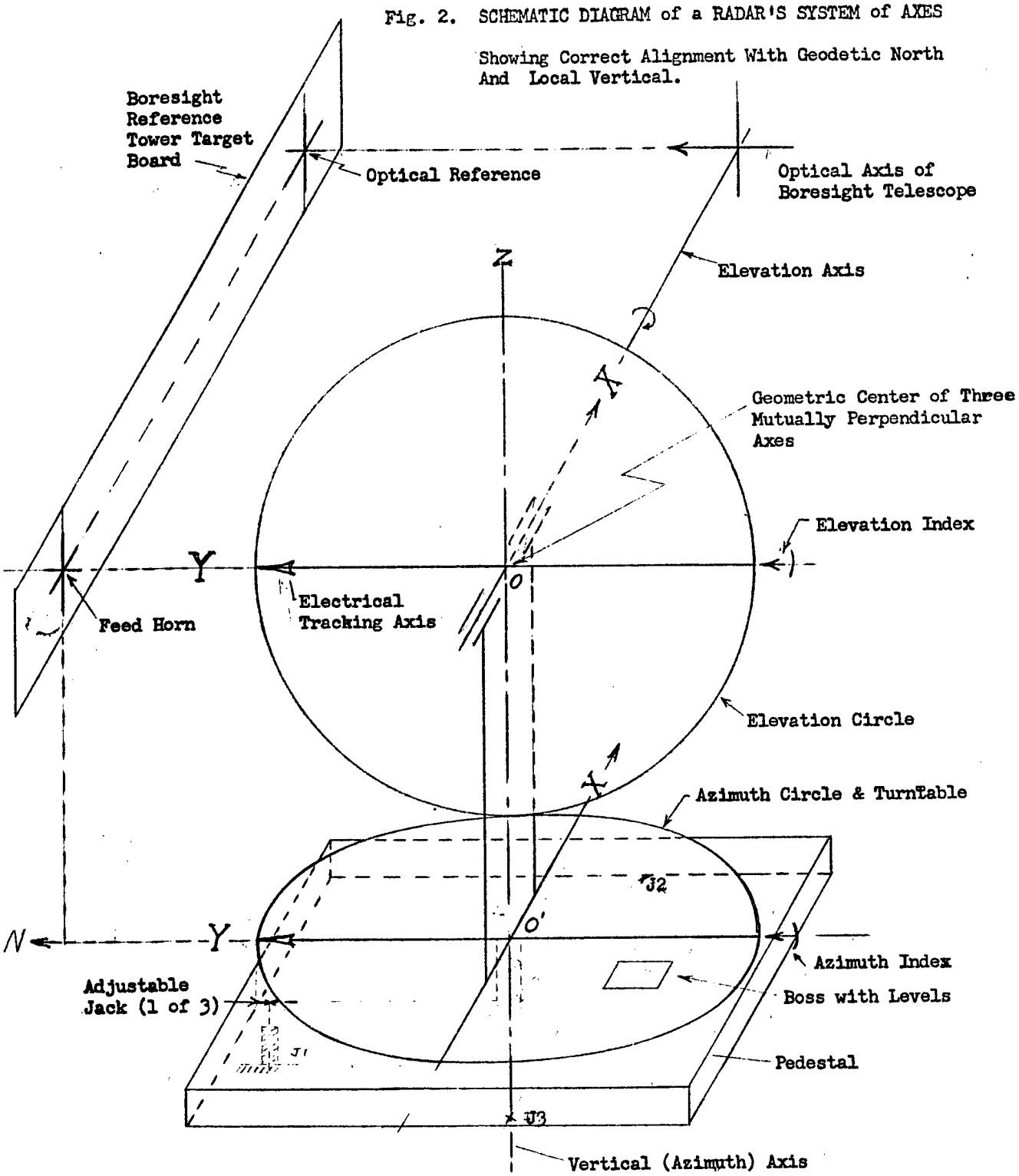


Fig. 3. ALIGNMENT ERRORS and their EFFECTS ON ΔA, ΔE in TERMS of RADAR ORIENTATION (A, E).

MISLEVEL

(1) $\Delta E = e_m \sin (A+\theta)$

(2) $\Delta E = \alpha \cos A + \beta \sin A$

(3) $\Delta A = \alpha \tan E \sin A - \beta \tan E \cos A$

STANDARDS ERROR ($\emptyset$)

(4) $\Delta A = b \tan E$ $\quad (b = \tan \emptyset)$

BENDING ERROR

(5) $\Delta E = d \cos E$ $\quad$ (or: $d \cos (E-E_m)$

ELECTRICAL AXIS DEVIATION (ℓ)

(6) $\Delta A = c \sec E$ $\quad (c = \ell$ in radians)

INDEX ERROR

(7) $\Delta A = g$

(8) $\Delta E = h$

ECCENTRICITY OF DIALS

(9) $\Delta A = (e/r) \sin (A - A_o)$

(10) $\Delta E = (e/r) \sin (E - E_o)$

Fig. 4. EFFECTS on Δ A of 1-mil ERRORS in MISLEVEL, ELECTRICAL AXIS DEVIATION, & STANDARDS ERROR