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

	Page
Foreword	i
Program	iii
The Method of Paired Comparisons By Dr. H. A. David	1
Measure of Competing Exponential Mortality Risks with Especial Reference to the Study of Smoking and Lung Cancer By Dr. Joseph Berkson	17
Army Research and Development By Dr. Richard Weiss	35
Prediction of the Reliability of Complex Systems By Dr. Nicholas E. Golovin	87
On the Repeated-Measurements Design in Biological Experiments By Ardie Lubin	123
Design of Experiments Using Germfree Animals By Stanley M. Levenson, Ole J. Malm, and Captain Richard E. Horowitz	133
The Development of Parameters for Determining the Resistance of Selected Missile Components to Microbiological Deterioration By C. Bruce Lee	151
Design of Environmental Experiments for Reliability Prediction By A. Bulfinch	171
Multidimensional Staircase Designs for Reliability Studies By David R. Howes	191
A Proposed Research Program for Providing a Quantitative Basis for Preventive Maintenance Policies on Ordnance Equipment By Walton M. Hancock and Randall E. Cline	199
Statistical Analysis of Various Parameters of Burning Characteristics of Flare Systems By Bossie Jackson	213
A Statistical Evaluation of the Pyrotechnic Electrostatis Sensitivity Tester By Everett D. Crane, Chester Smith, Alonzo Bulfinch	239

TABLE OF CONTENTS (Cont'd)

Page

Dispersion Strengthening Analysis of Cermets
By John M. Woulbroun* . . . . . .

Experimental Determination of "Best" Component Levels in Thermal Power Supplies (U)
By Sheldon G. Levin 263

Medical Health Statistics
By Dr. Wilford J. Dixon 265

Sampling in Biological Populations
By Dr. D. B. DeLury 277

The Application of Fractional Factorials in Missile Test Programs
By Paul C. Cox 285

The Design and Re-design of an Experiment
By C. W. Mullis 291

Estimating the Parameters of a Modified Poisson Distribution
By A. C. Cohen 303

Detecting and Quantifying Guess Responses in the Rating of Statements by a Method of Successive Intervals
By Lee E. Paul 309

Design for Estimation by Covariance Techniques
By Morris Rhian 317

Design of an Experiment to Evaluate a Bio-assay with Non-parallel Slopes
By Albert L. Fernelius 327

The ORO Aircraft Vulnerability Experiment
By Charles A. Bruce and Bruce Taylor 333

Operational Hit Probabilities of Experimental Antitank Weapons
By J. D. Reed, R. E. Tiller, and J. P. Young 343

Elimination of Bias Introduced by Transformation of Variables
By Jerzy Neyman and Elizabeth L. Scott 353

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

TABLE OF CONTENTS (Cont'd)

Page

Mathematical and Statistical Principles Underlying Chemical Corps Inspection Procedures for Product Verification
By Henry Ellner and Joseph Mandelson 373

Measuring a Complex Field Operation
By K. L. Yudowitch 395

The Conduct of Military Field Research on a Shoe-String
By A. J. Eckles, III 403

Sample Order Statistics of the Circular Normal Distribution
By Helen J. Coon* . . . . . .

Determination of Systematic Errors in Tracking Radar
By Victor B. Kovac** 417

---

* 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

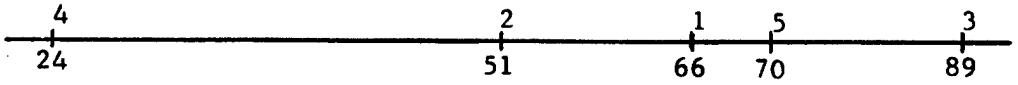
0.644, 6.62	for the Round Robin tournament,
0.656, 6.56	for the Knock-out tournament,
0.686, 6.43	for the Double Elimination tournament,
0.706, 7.08	for Maurice's tournament.

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

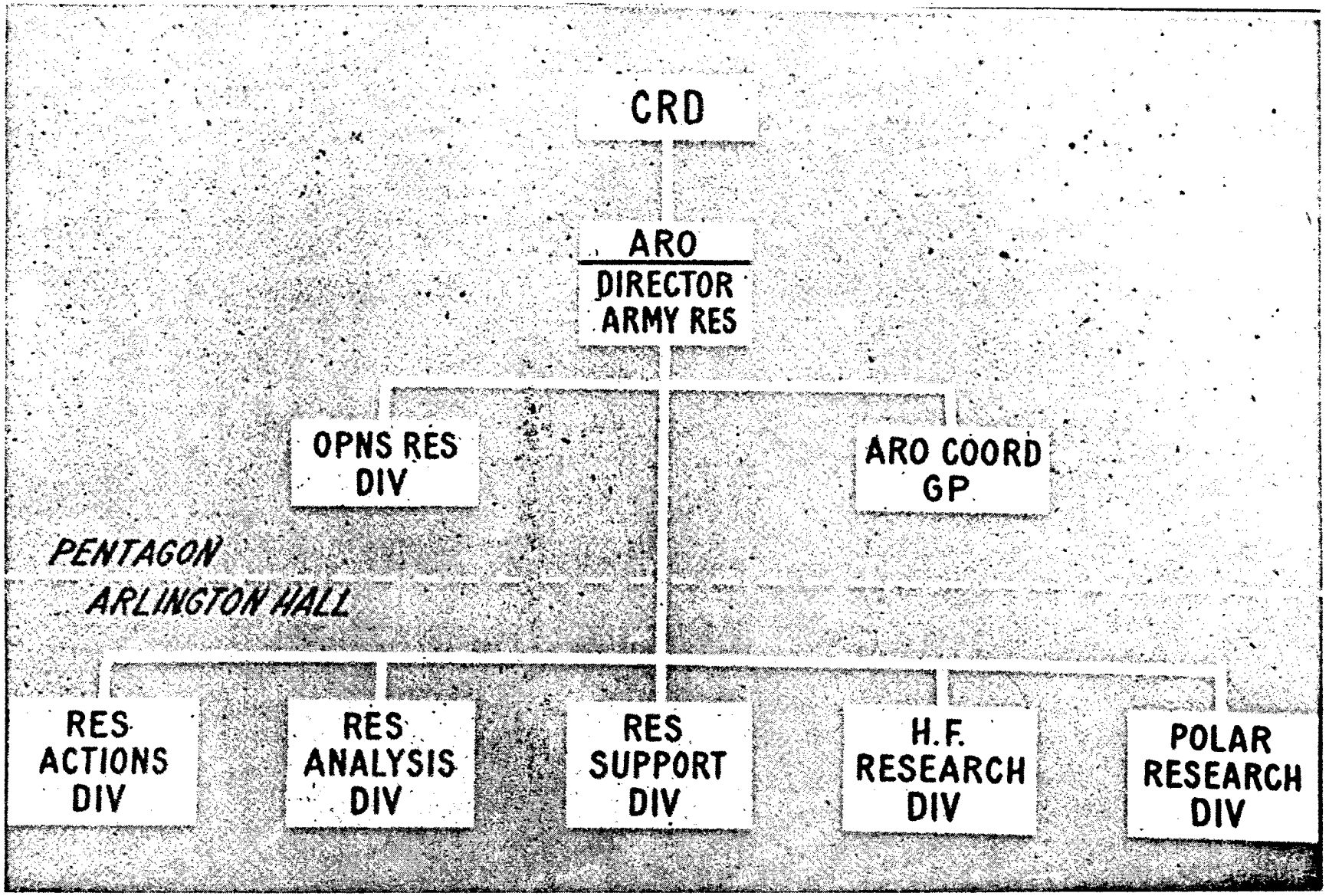
Richard A. Weiss
Army Research Office

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

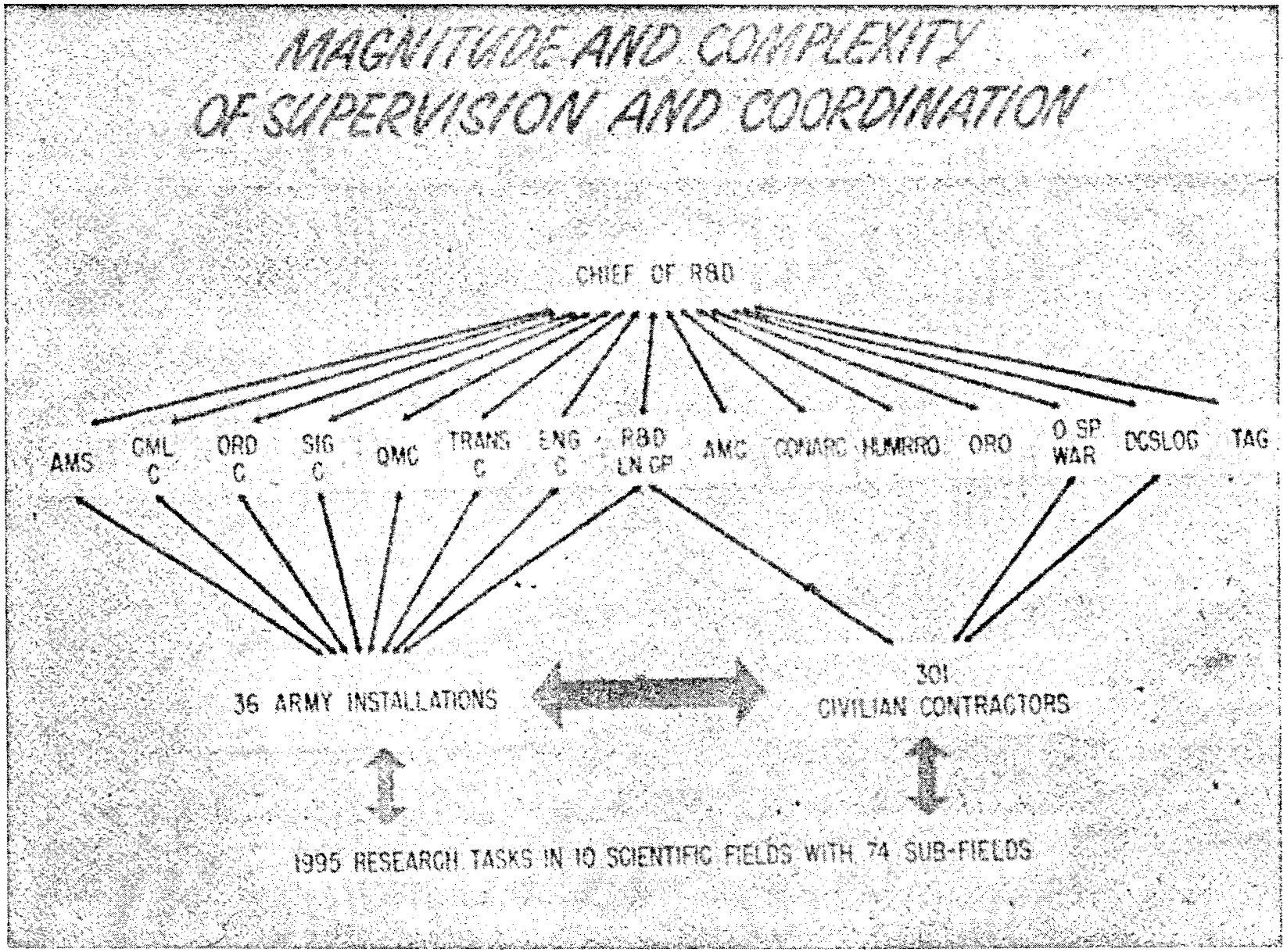


Slide 1

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

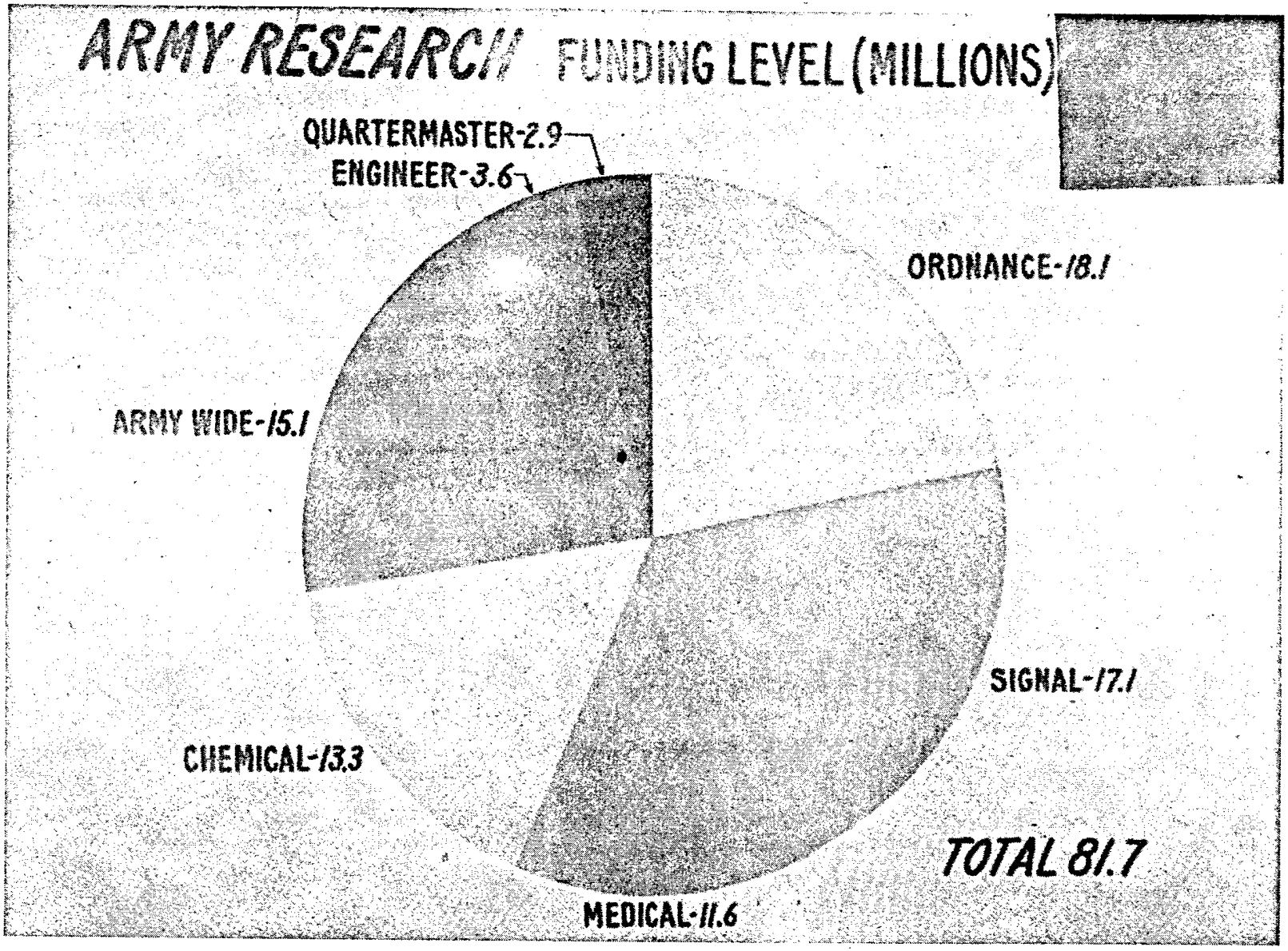
I don't know whether this can be seen, but this is an idea of the magnitude of supervision and coordination that the Chief of Research and Development has. The first seven blocks on the left are the seven Technical Services - Army Medical Service, Chemical Corps, Ordnance Corps, Signal Corps, Quartermaster 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)

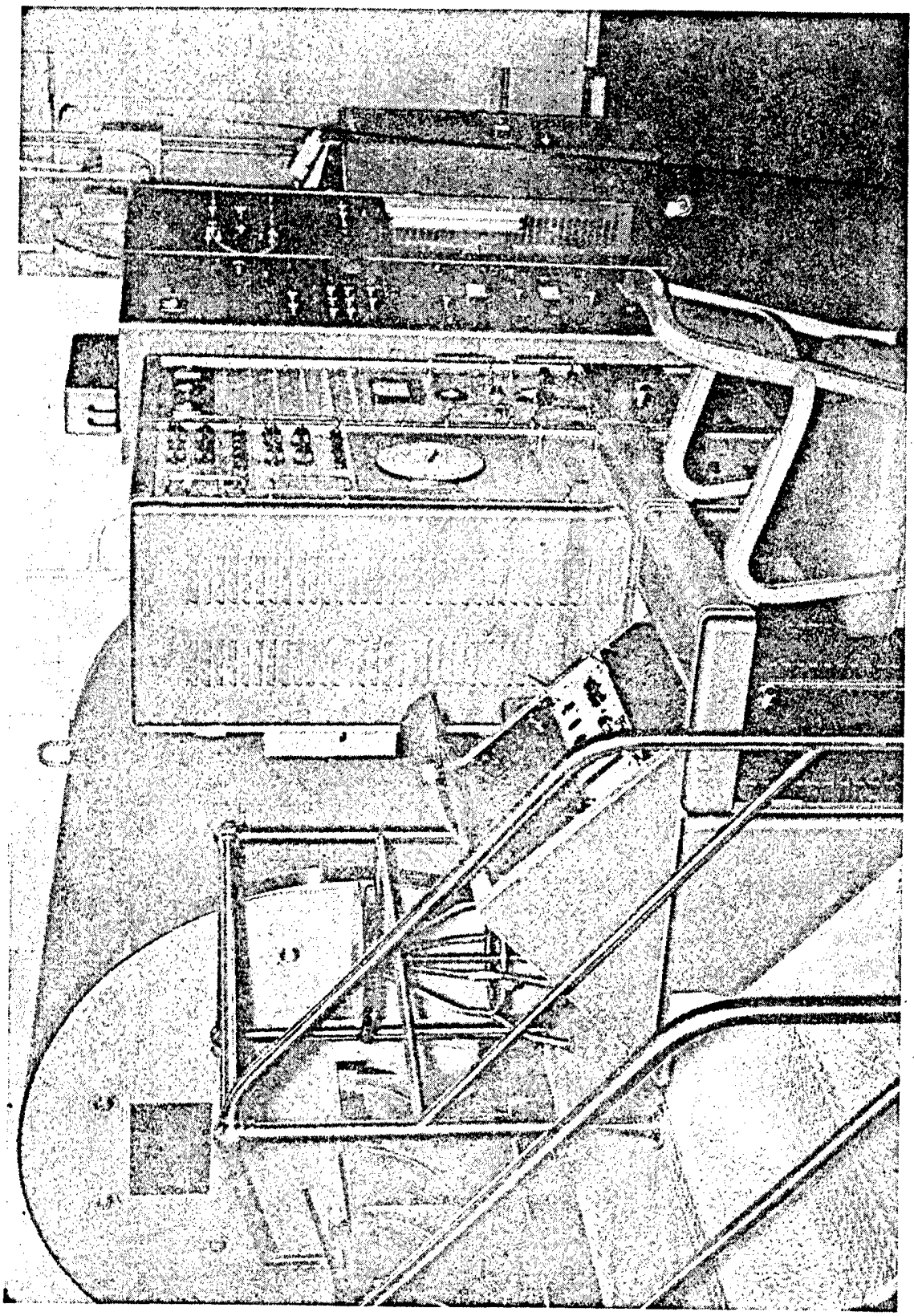


Slide 3

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

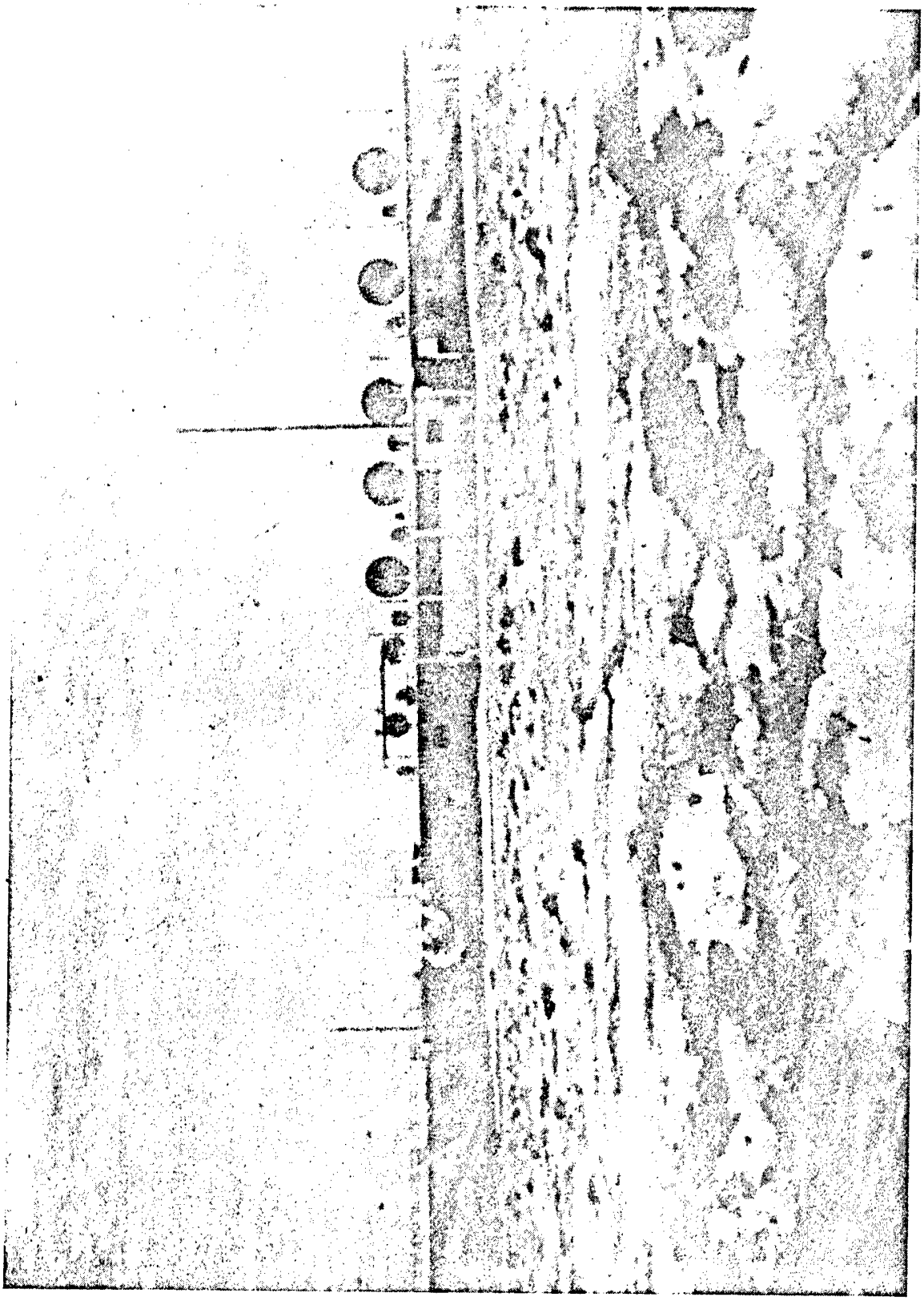
This is a slide of the whole body radiation counter at the Walter Reed Army Institute for Medical Research. It is possible to put a man inside so that the radiation emitted by virtue of any radioactive material that is in his body can be counted by a bank of scintillation counters. As a matter of fact, 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.

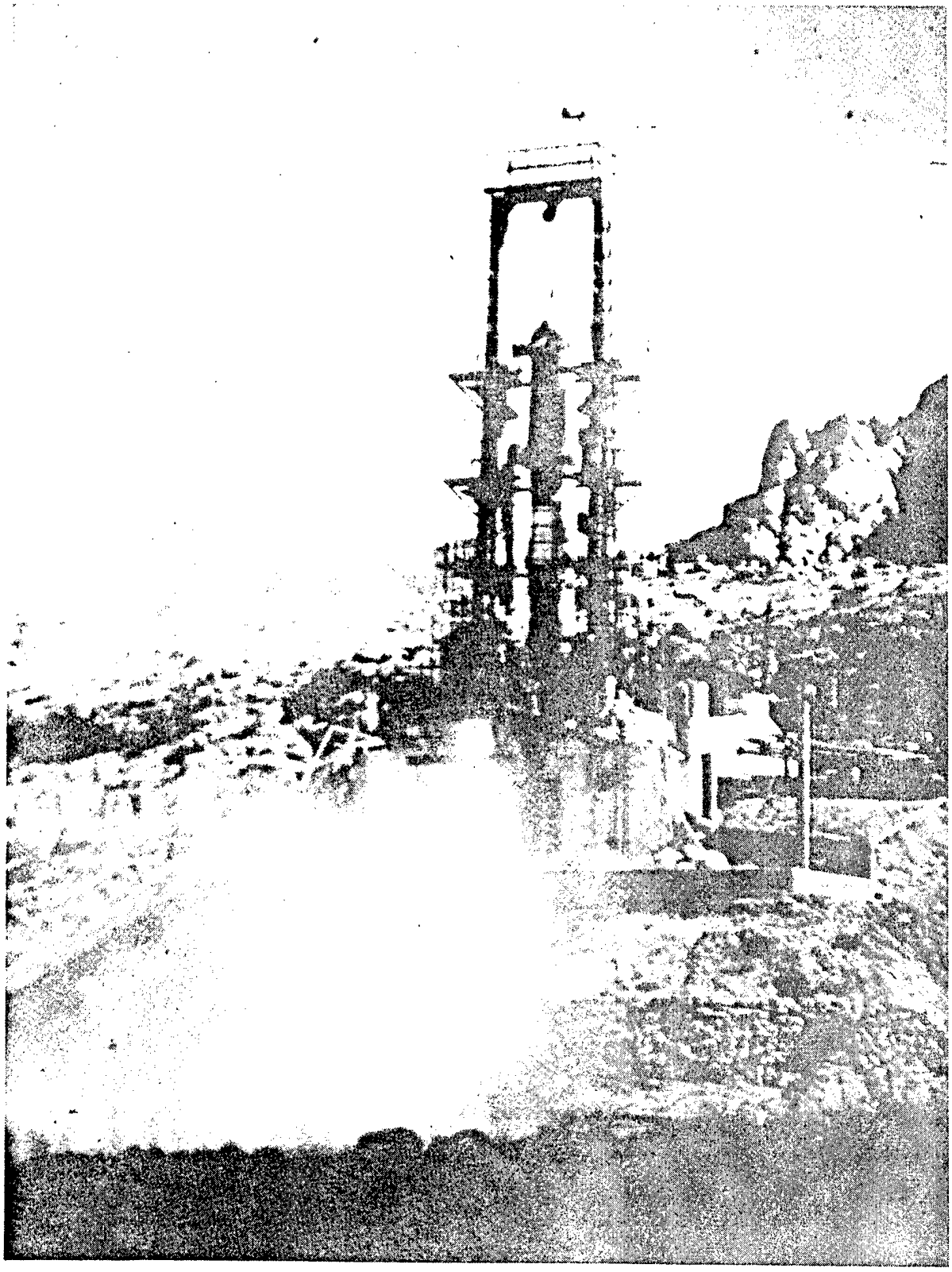


Slide 5

White Sands Proving Ground

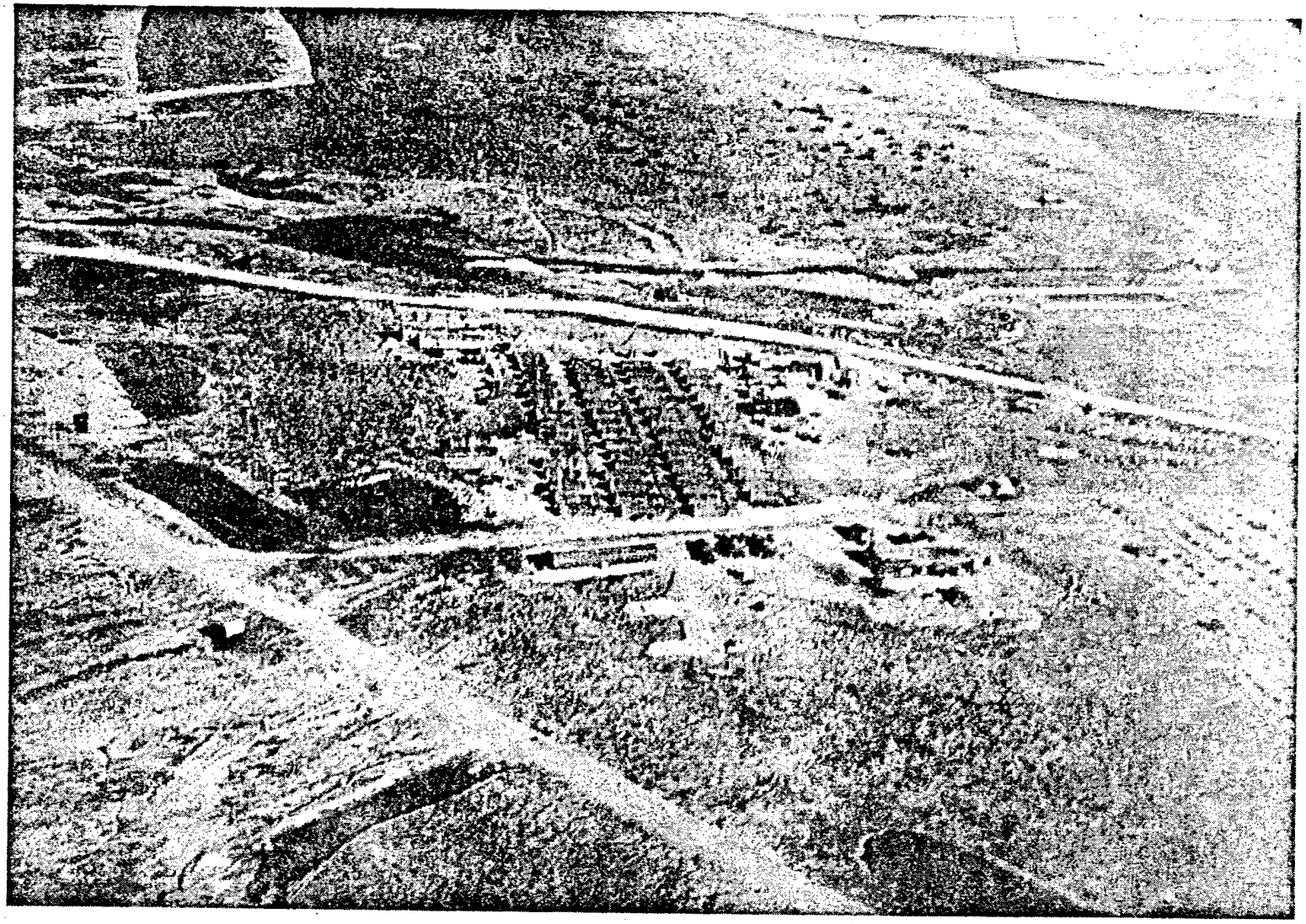
Slide 6

Engine on Static Test Stand



Slide 7

Thule, Greenland



Slide 7

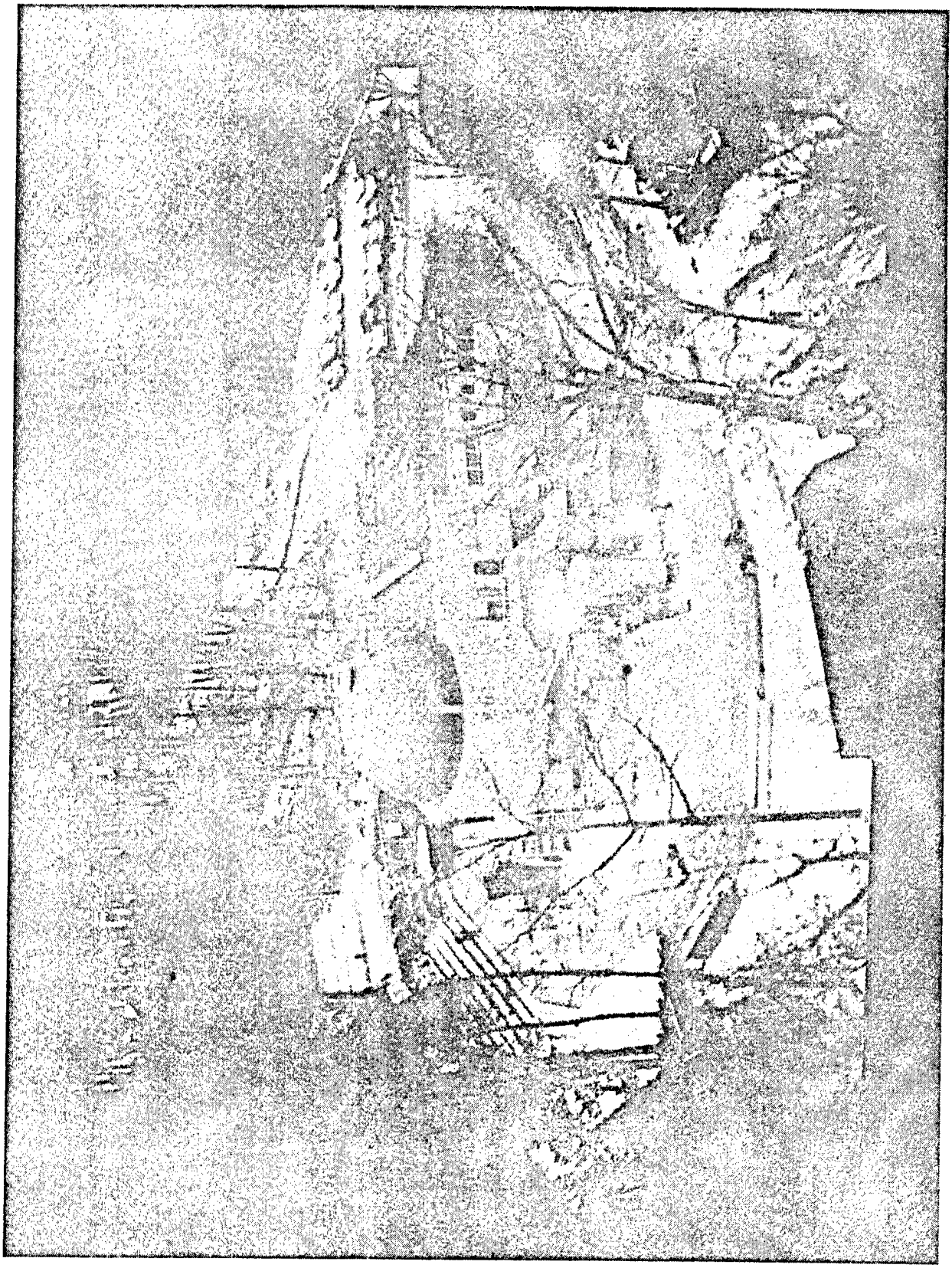
The following slide was taken in Thule, Greenland. Here there are something like nine disciplines covered in the 50 research tasks that are being undertaken by all the Technical Services. This is the area where the Engineers and the Transportation Corps, particularly, are engaged in a major program in 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.

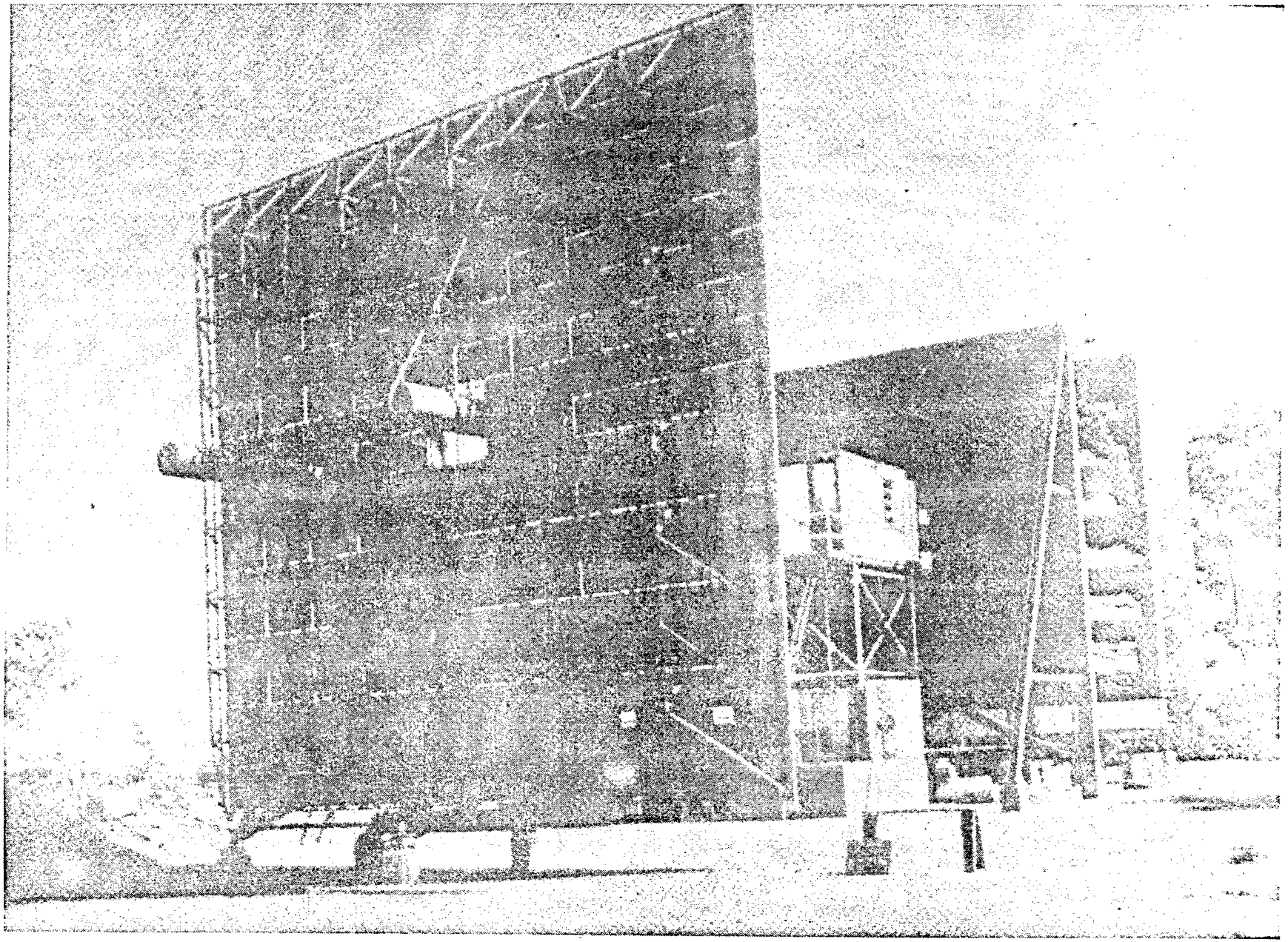


Slide 8

Power Reactor, Fort Belvoir

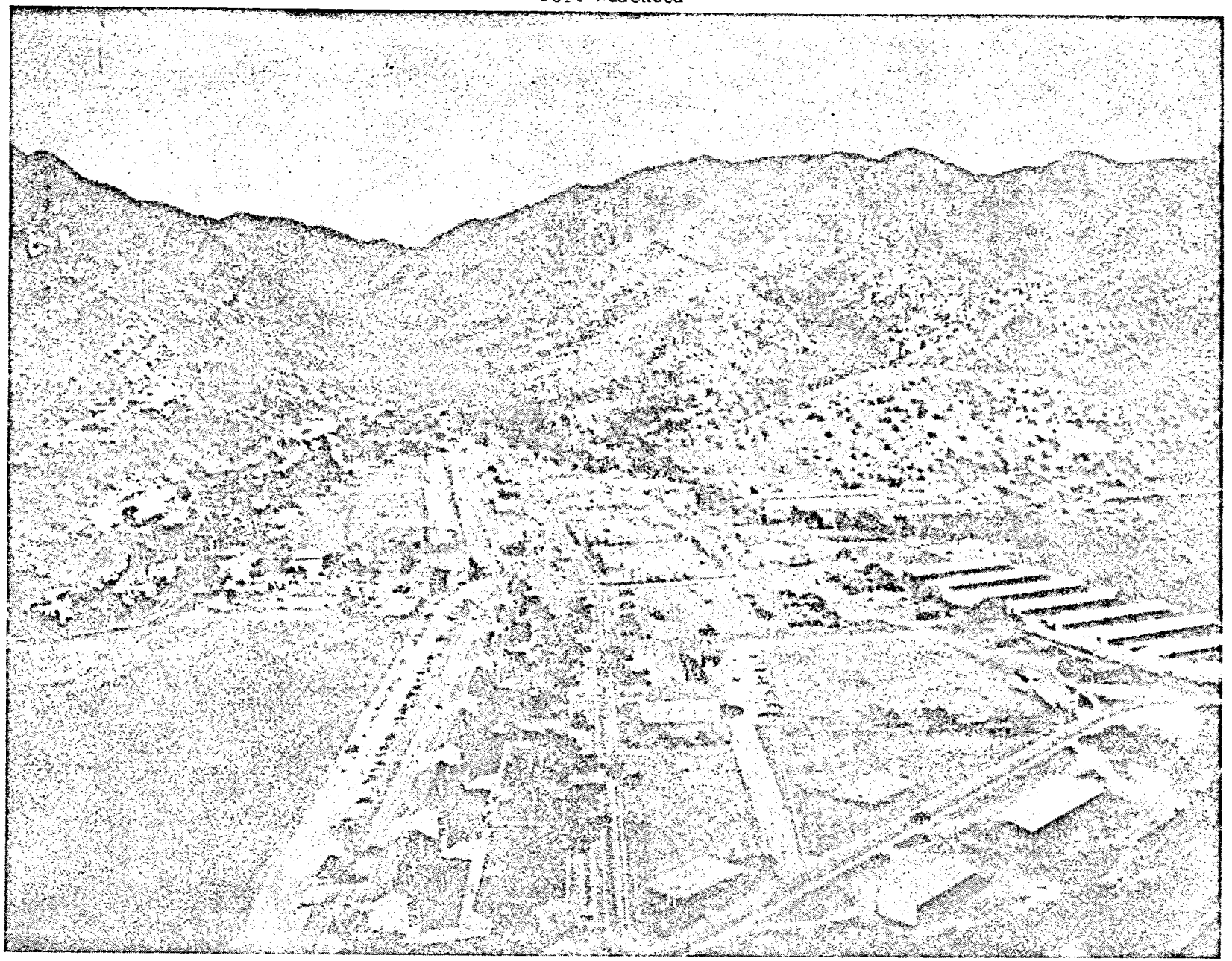
Slide 9

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.

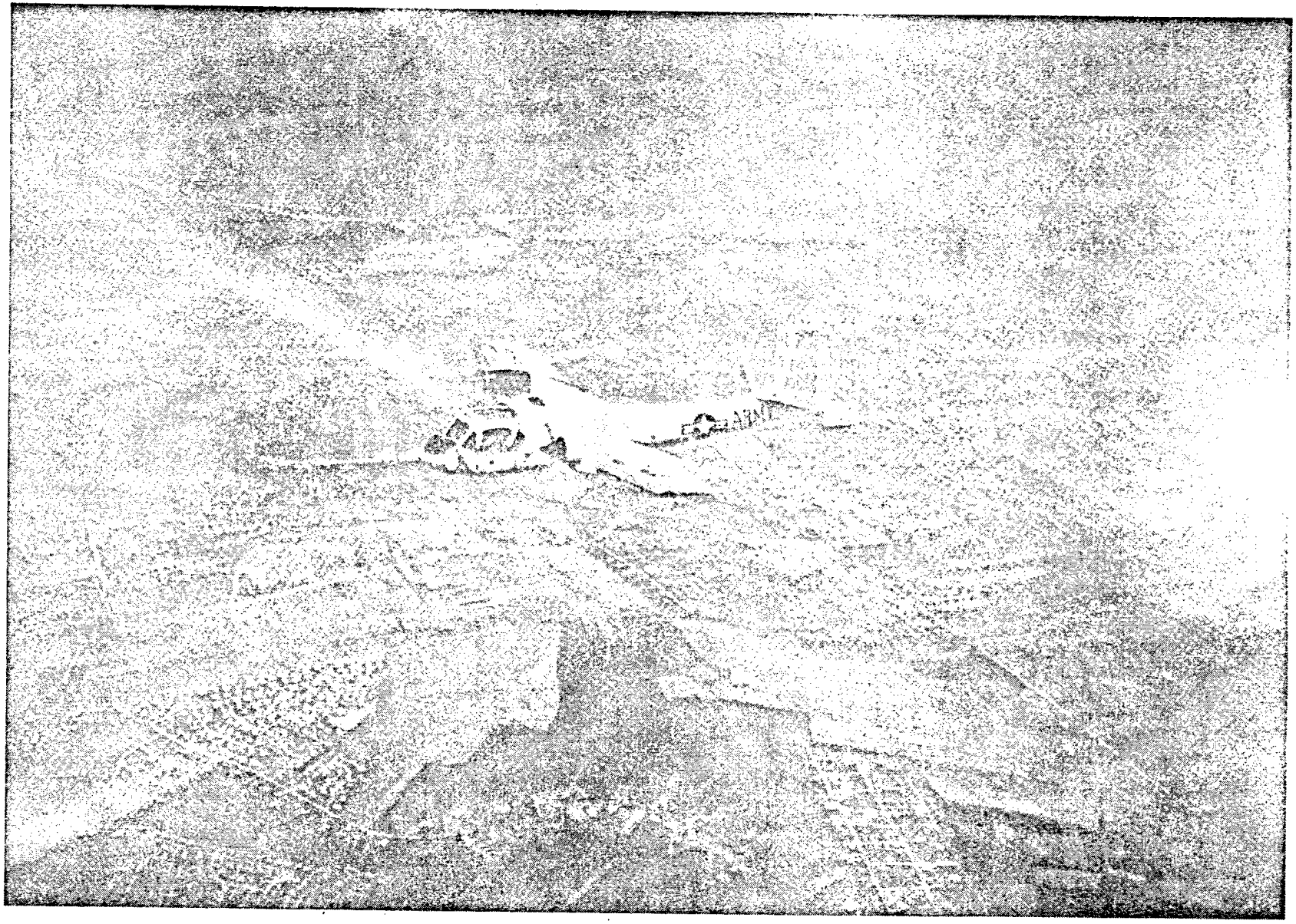
If the Communists ever use chemicals against us, we must be prepared to meet such an attack. Chemical Corps is working on this. On the other hand, Chemical Corps has performed recent tests which prove that nonlethal gases can incapacitate without killing, leaving no harmful after effects on humans or 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

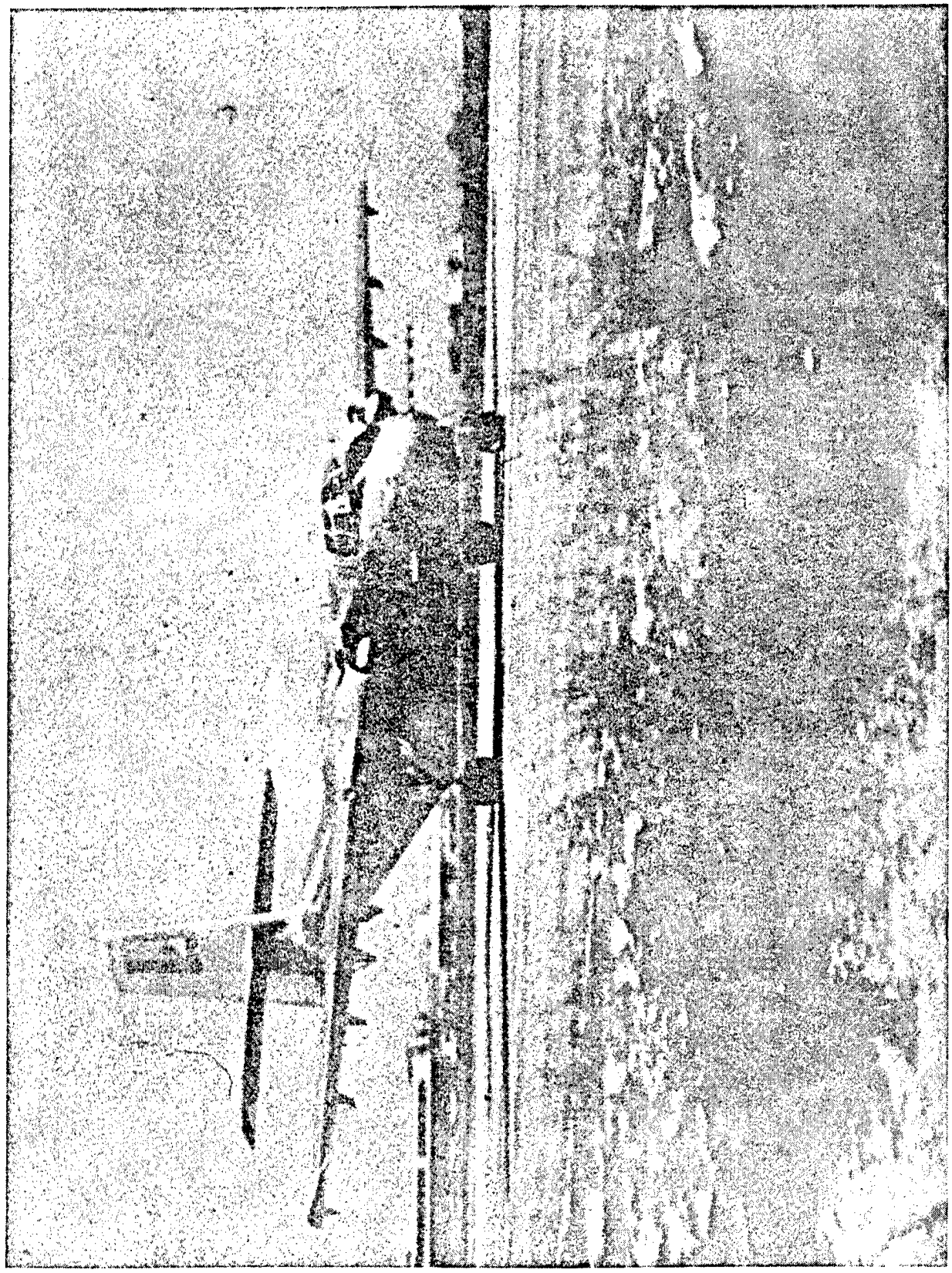
The first slide is the Army's Mohawk - reconnaissance and observation aircraft. This is a high-speed, short-take-off and landing aircraft for visual, photographic and electronic observations for shallow penetration into the enemy lines in the order of 25 to 40 miles.

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

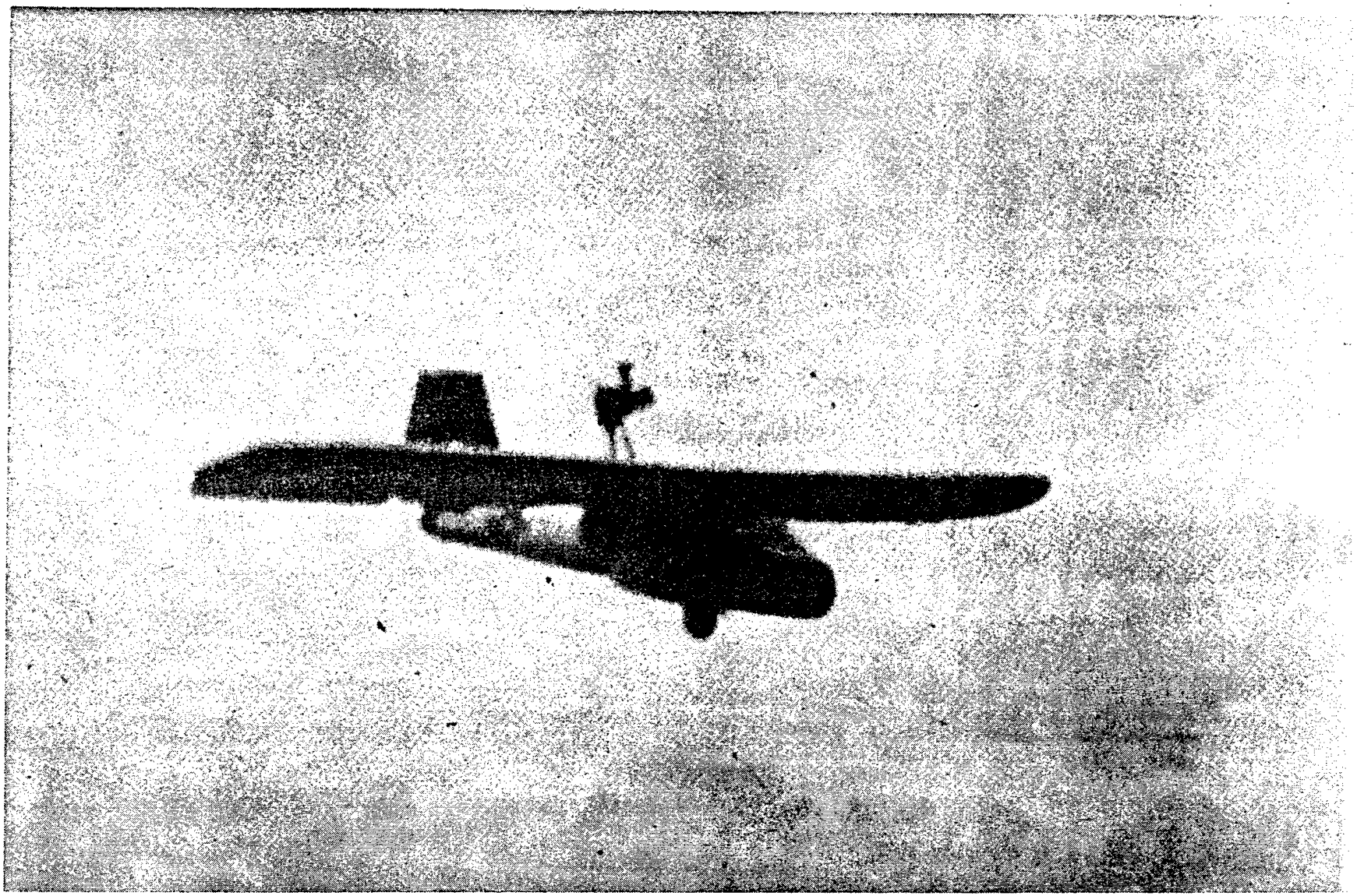
The next slide indicates work done by Goodyear on the Inflatoplane. This is a compact, rubberized craft which can be inflated by means of high-pressure CO_2. Parachuted out of aircraft, it can be assembled on the ground and flown from crude landing fields. Here it is, assembled and in the process of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

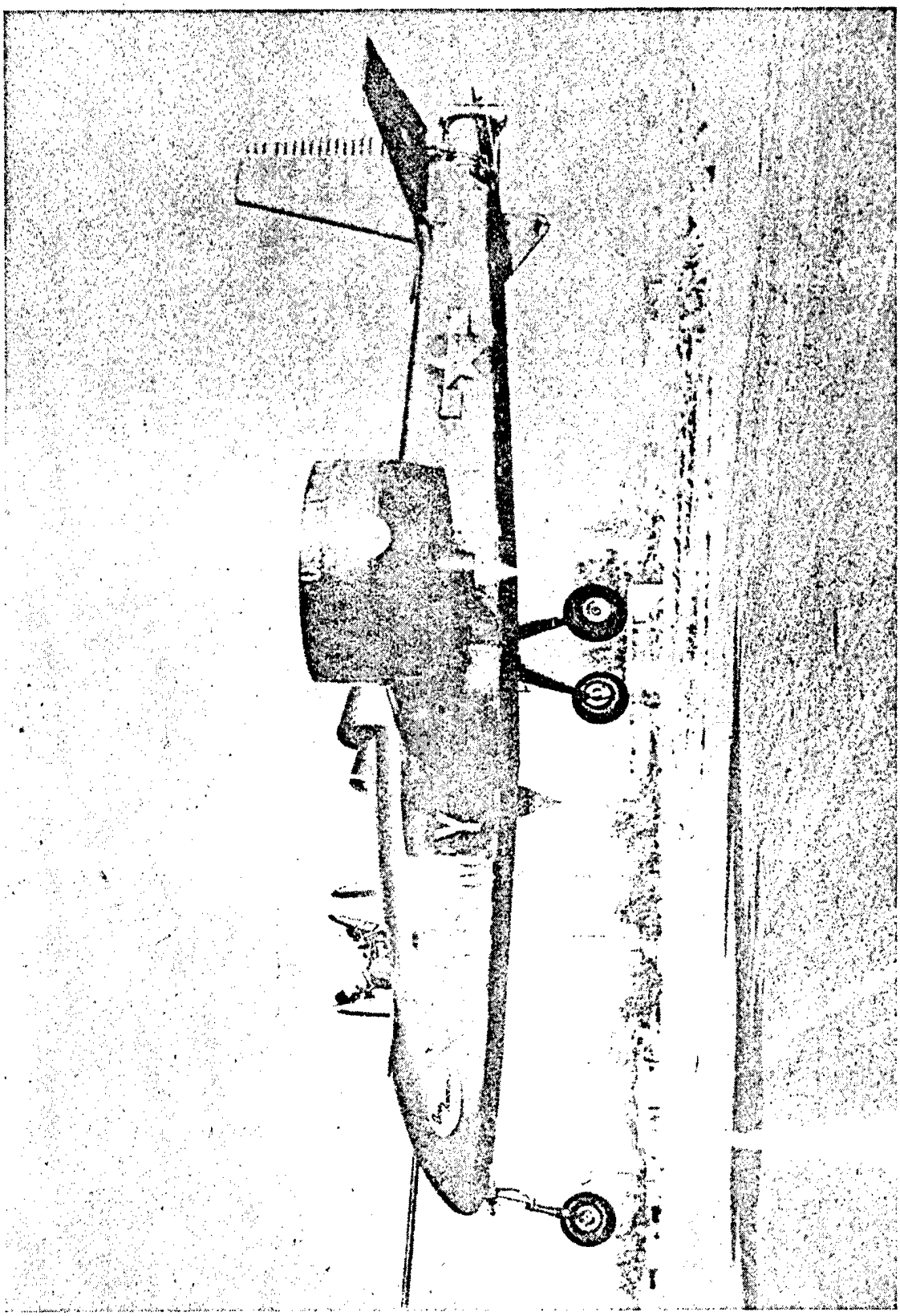


Slide 12
Caribou

Slide 13

Inflataplane





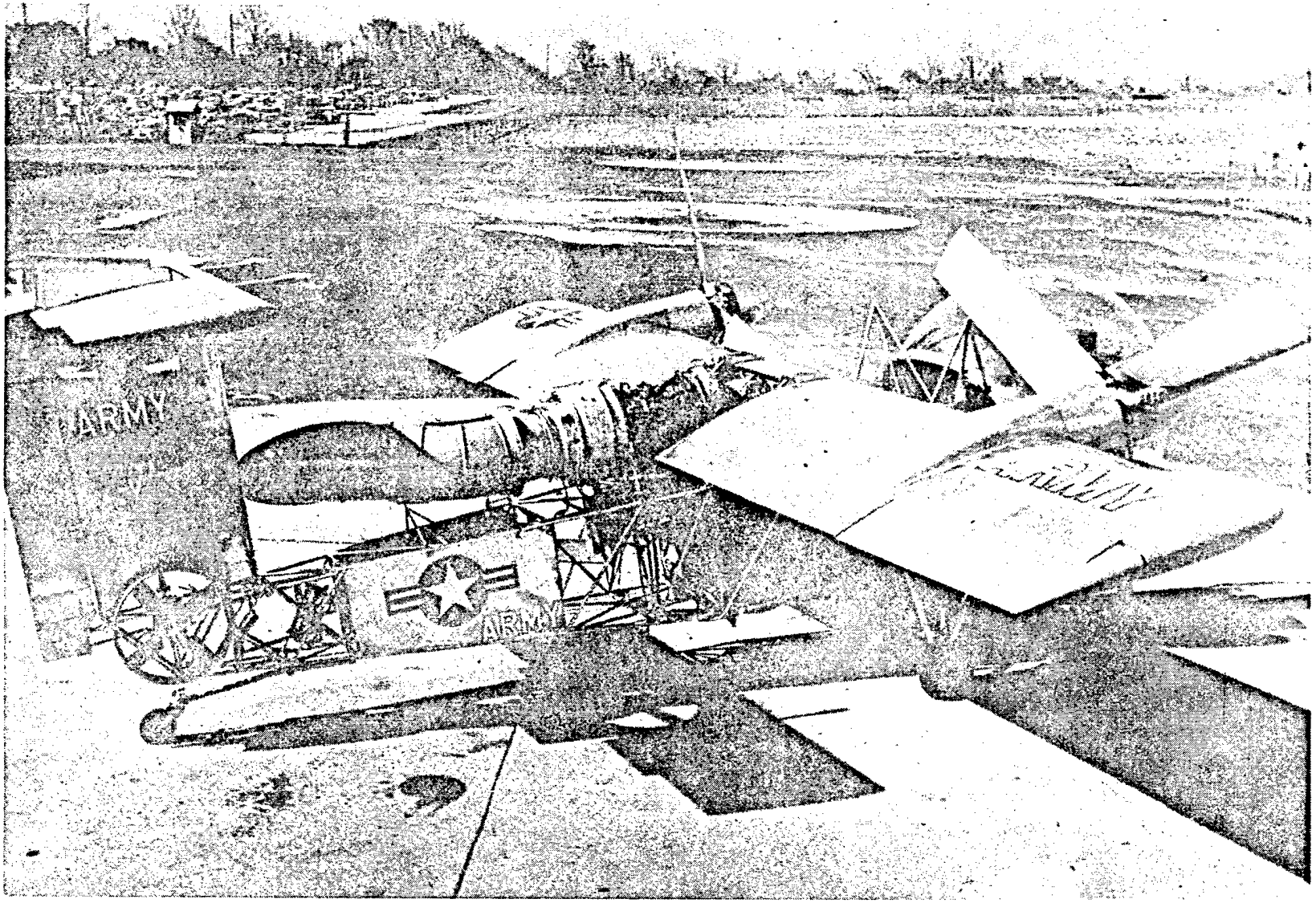
Slide 14
Ducted Fan Doak

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.

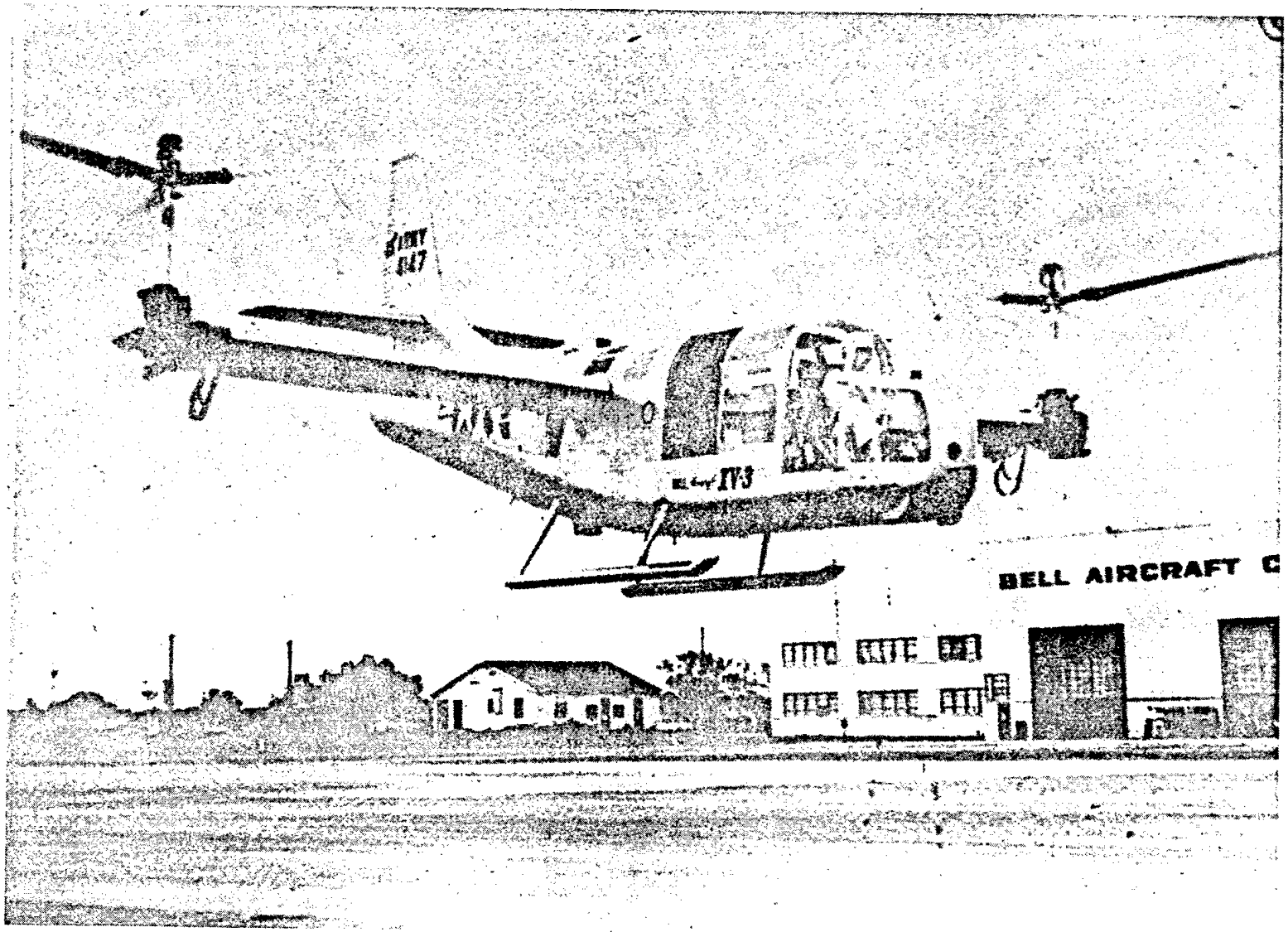
Slides 16, 17

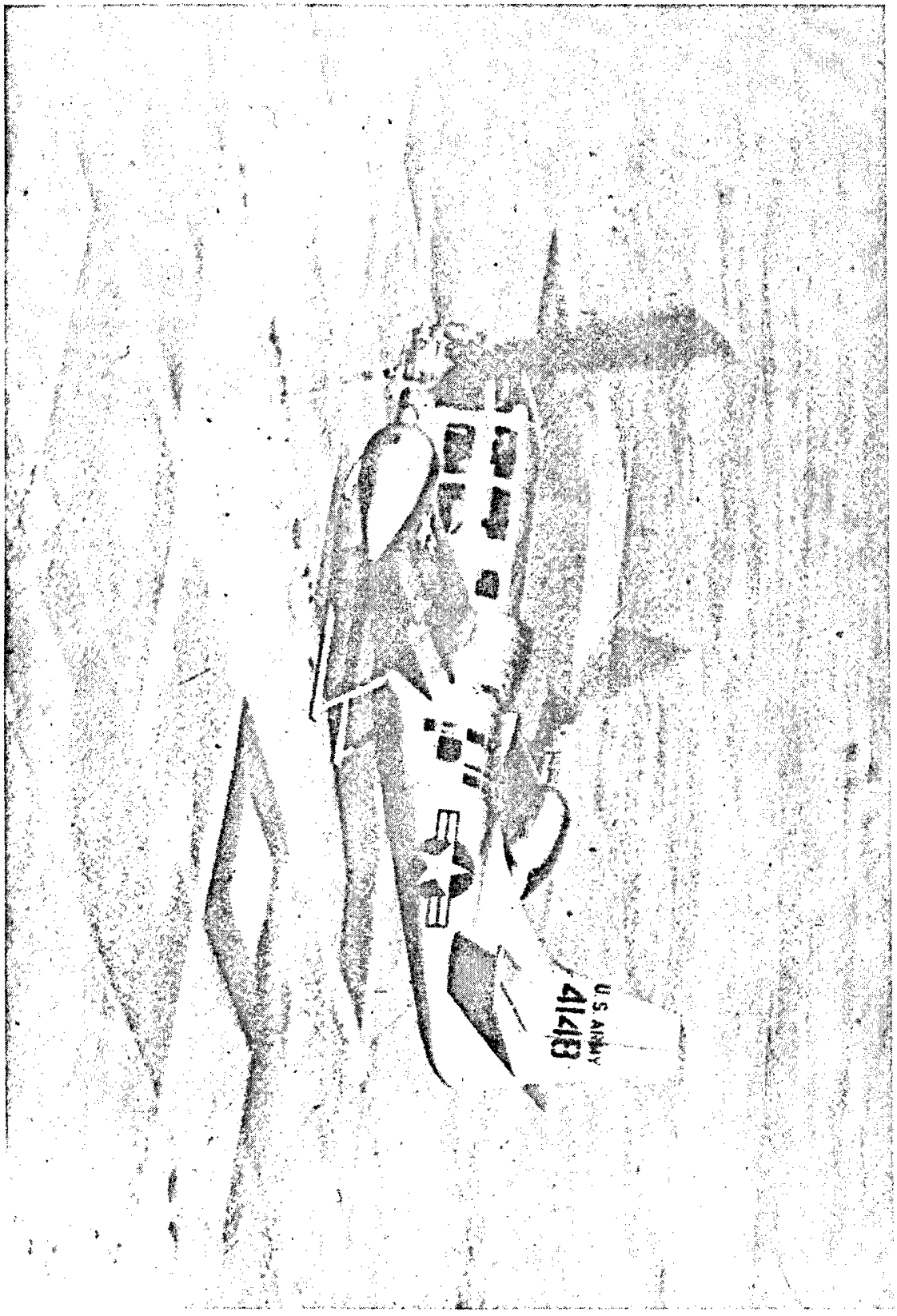
In the converti-plane made by Bell Aircraft you can see, in the next slide, the lift phase, and in the next, the forward flight phase where the propeller has turned through 90 degrees. They have successfully gone through the transition phases of several of these experimental aircraft.

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

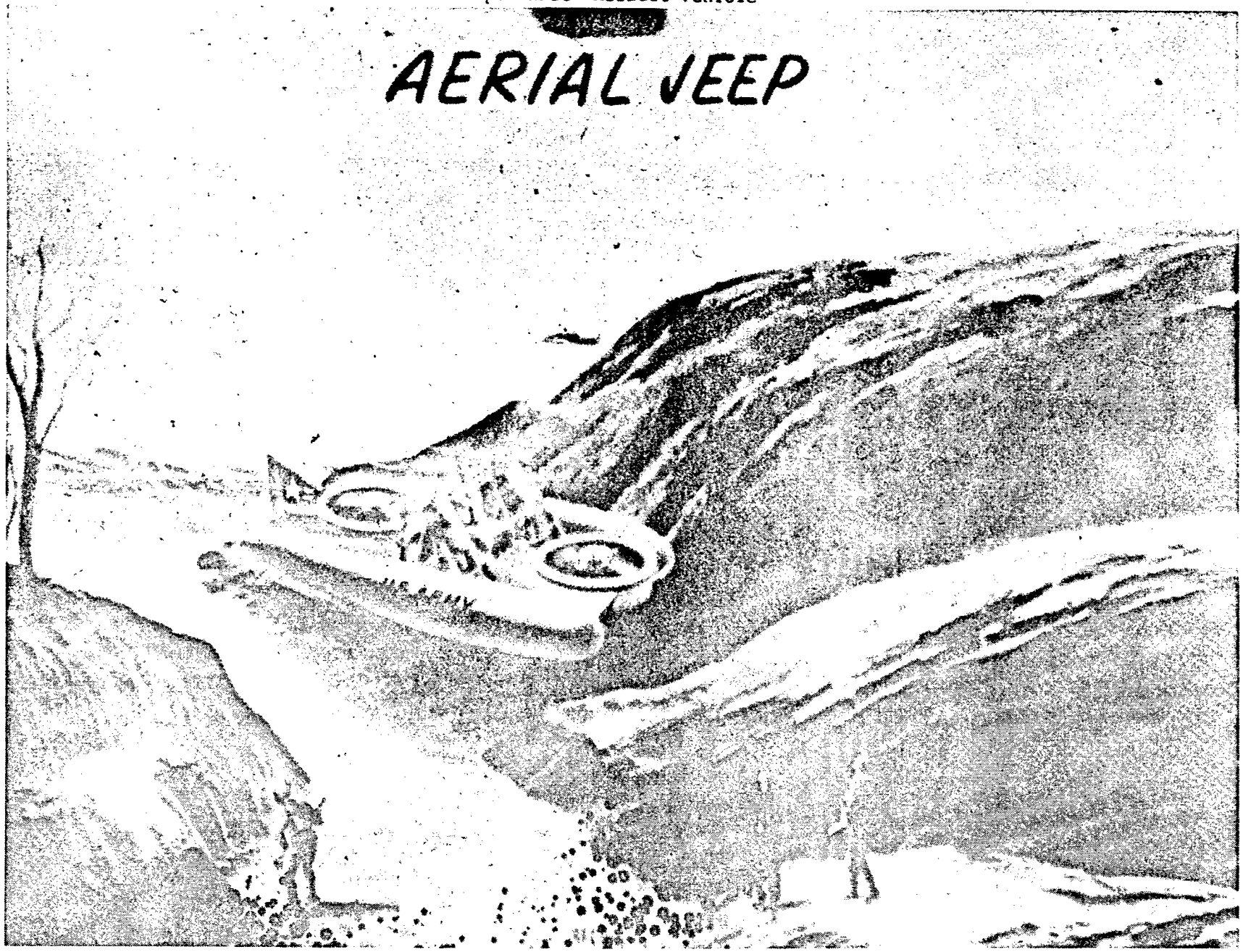




Slide 17

Bell Convertiplane - Forward Phase

Concept Phase Assault Vehicle



Slide 18

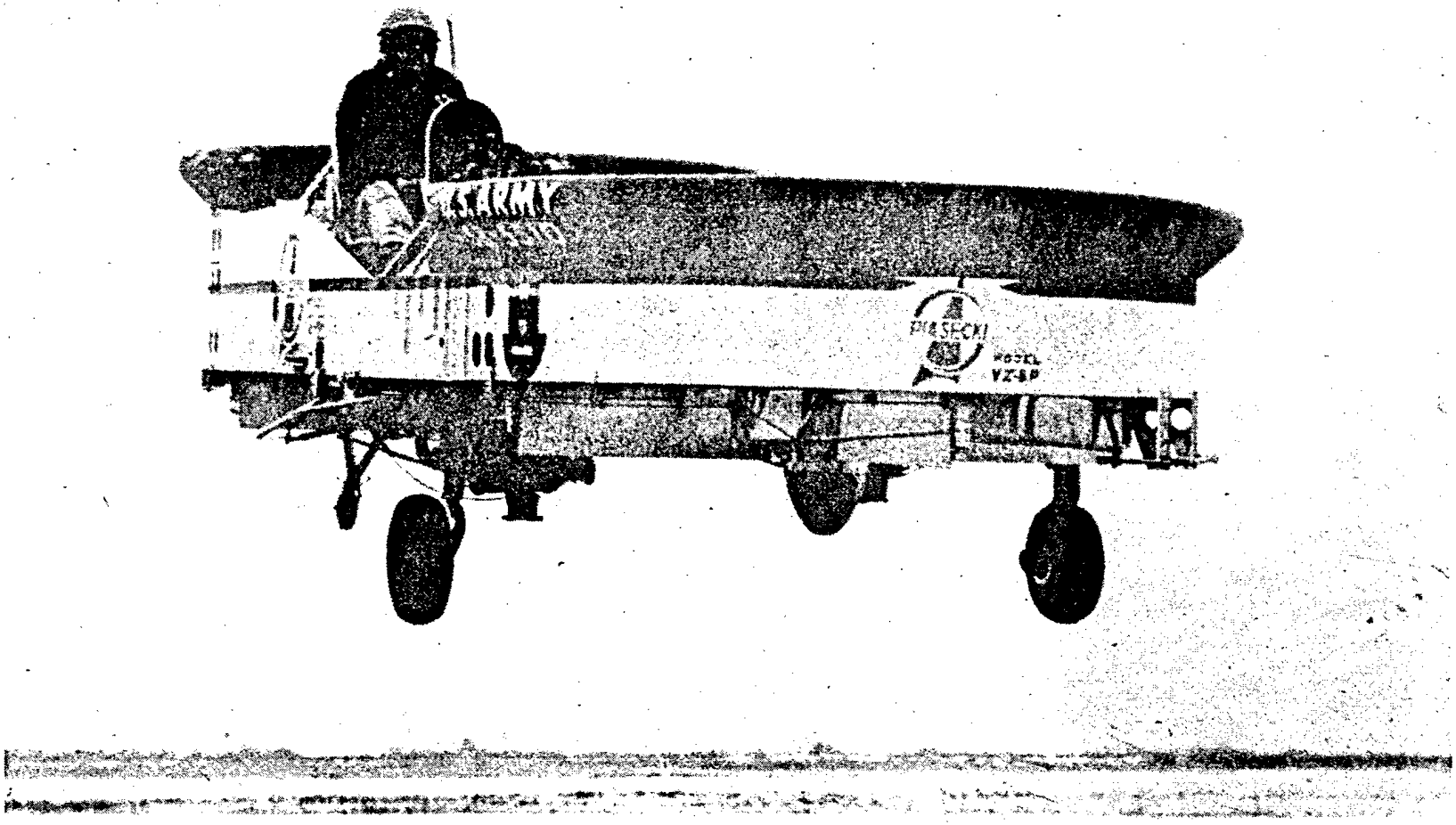
First is a Piasecki prototype aerial vehicle we used to call the "Jeep," however that is a patented name now, so we have to call it something else. It's just a flying vehicle - an artist's concept, as a matter of fact.

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.

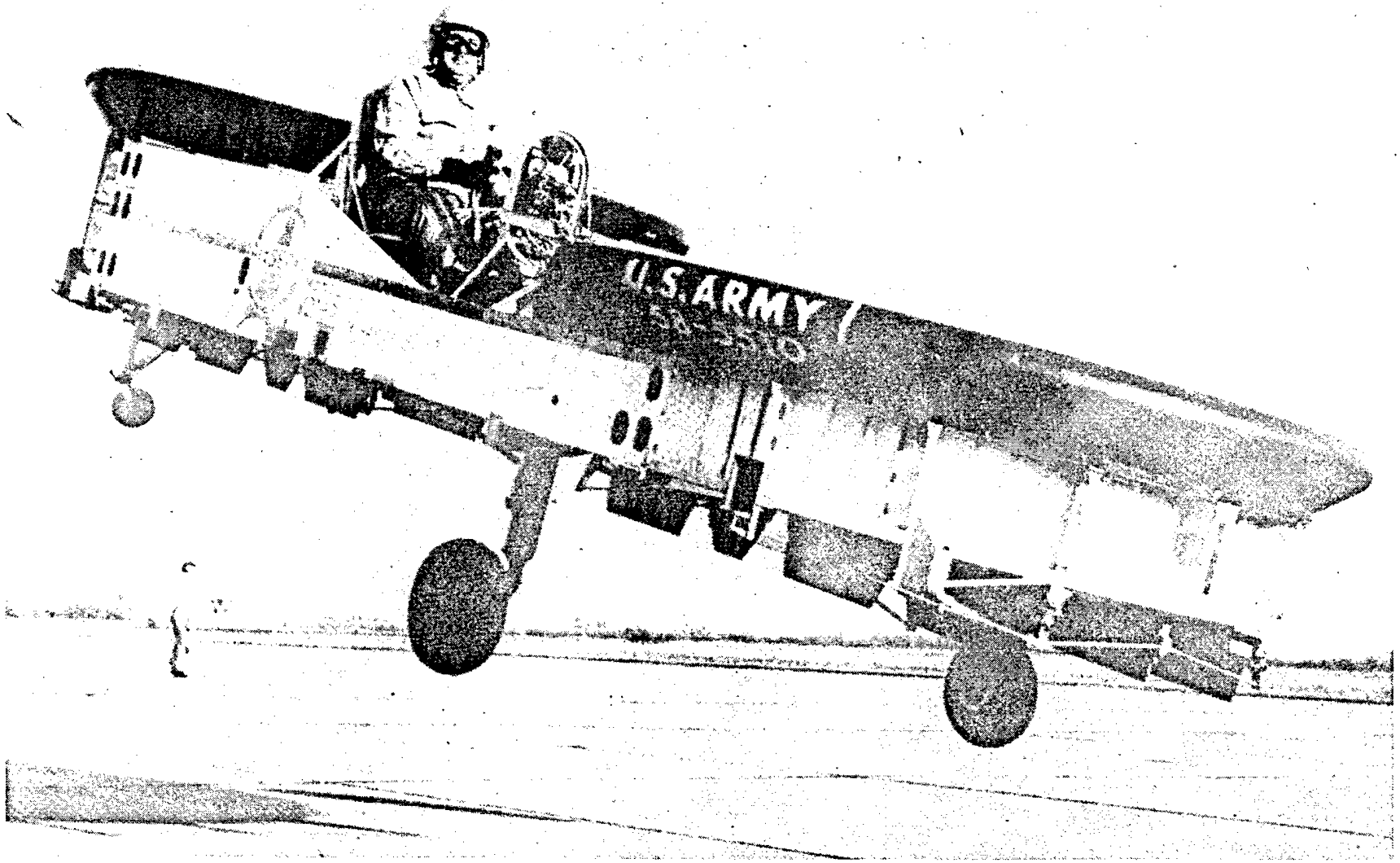
Slide 19

Aerial Jeep - Flight



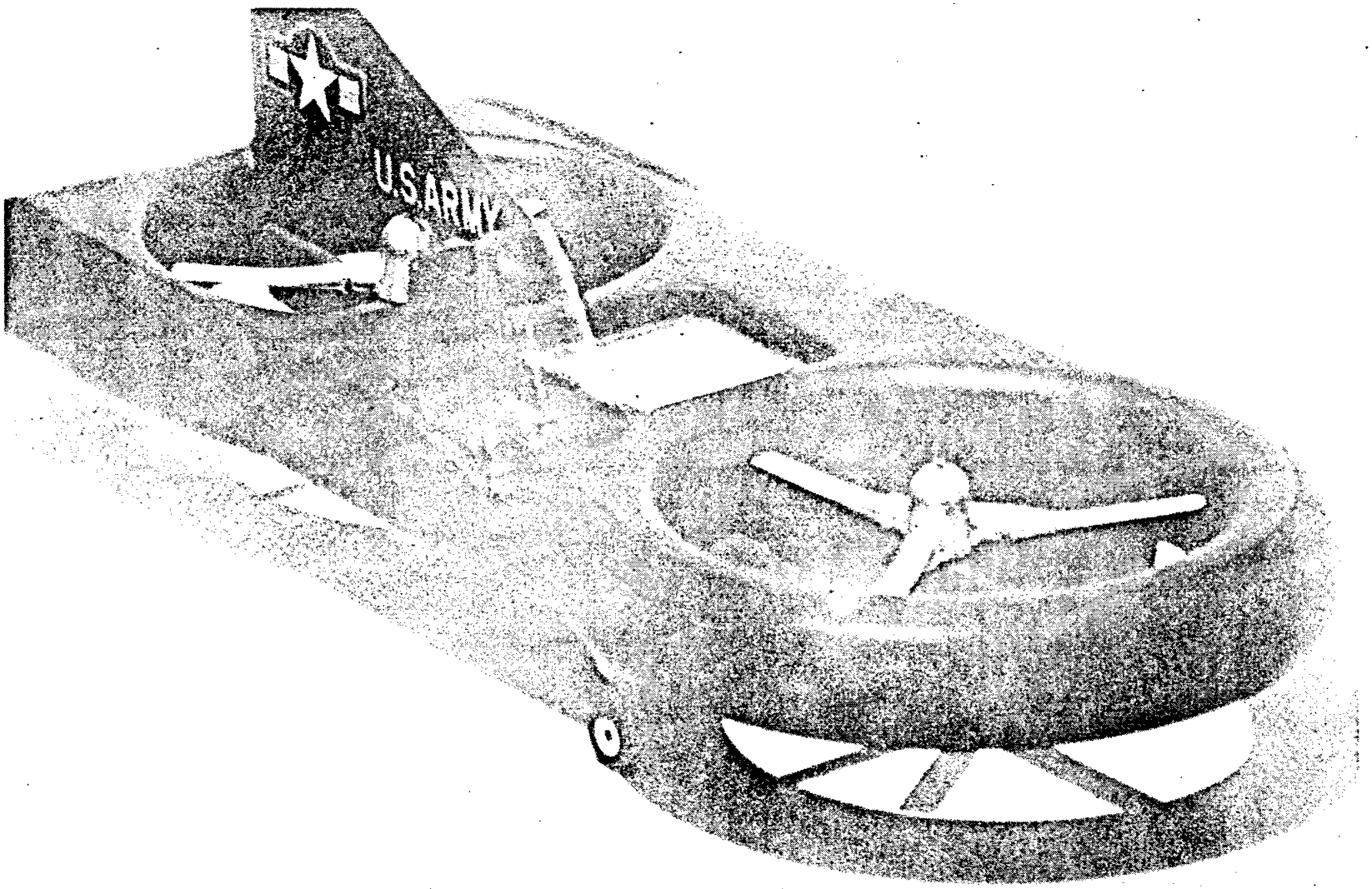
Slide 20

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.

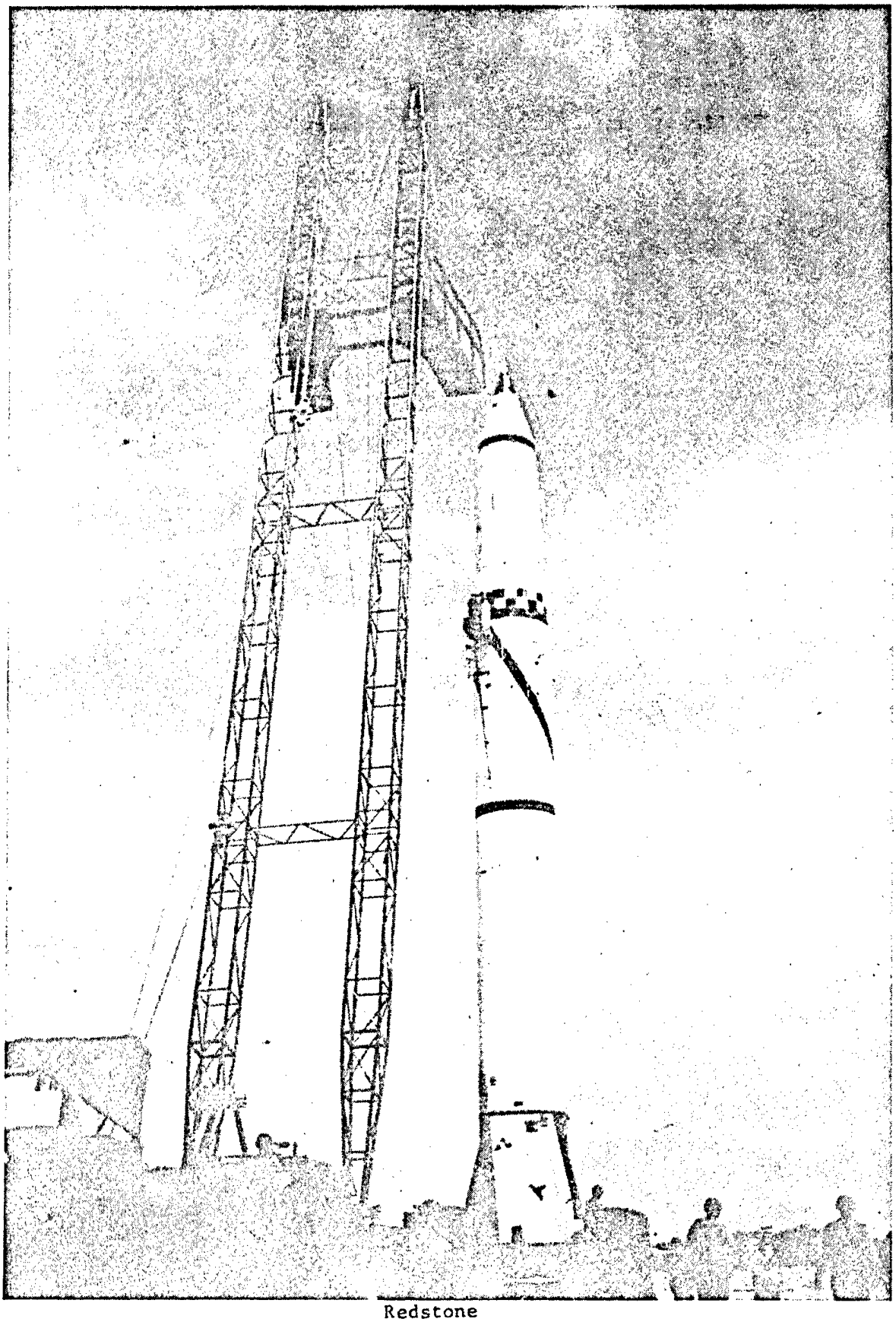
Slide 22

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.

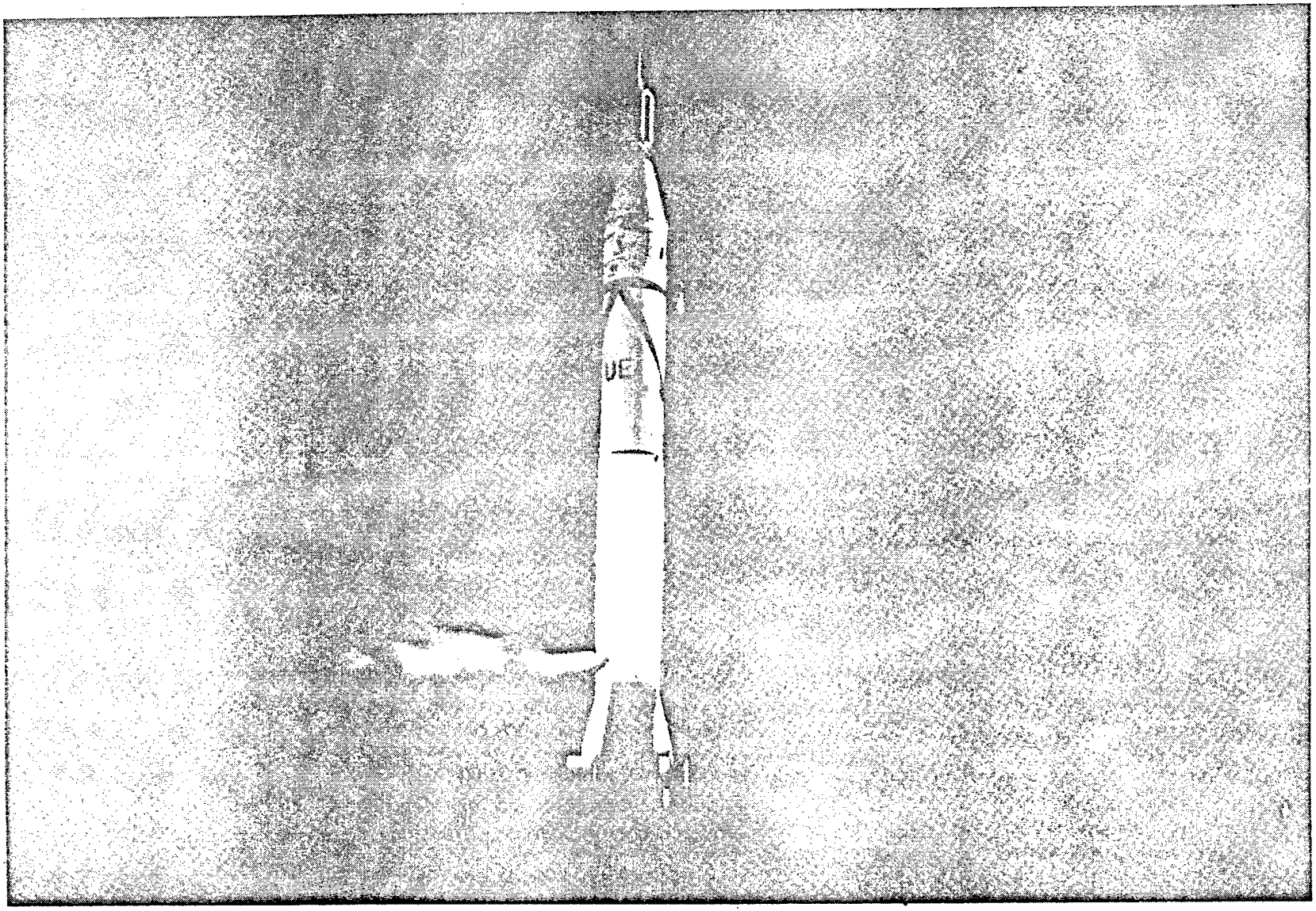
Slide 22



Redstone

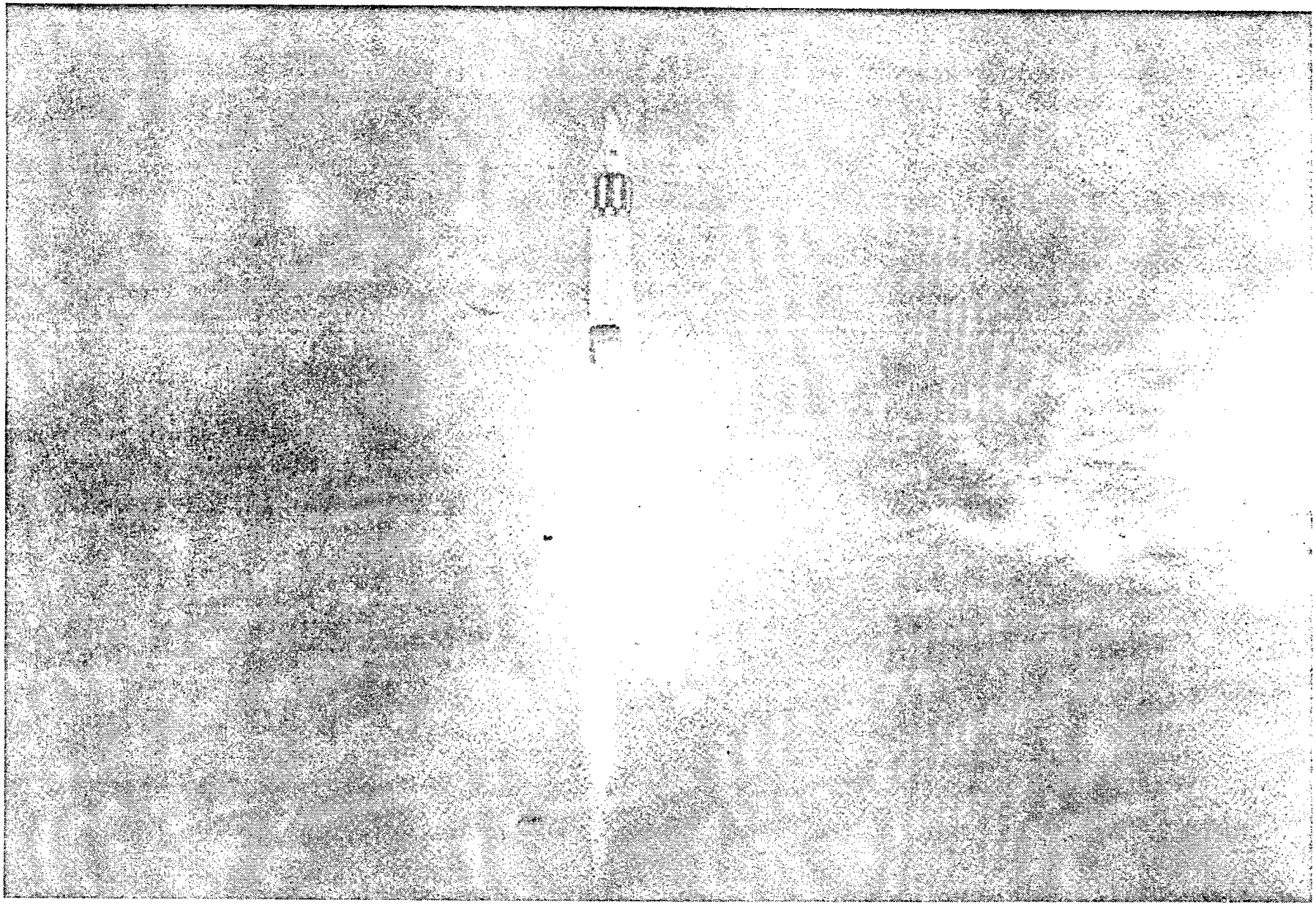
Slide 23

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

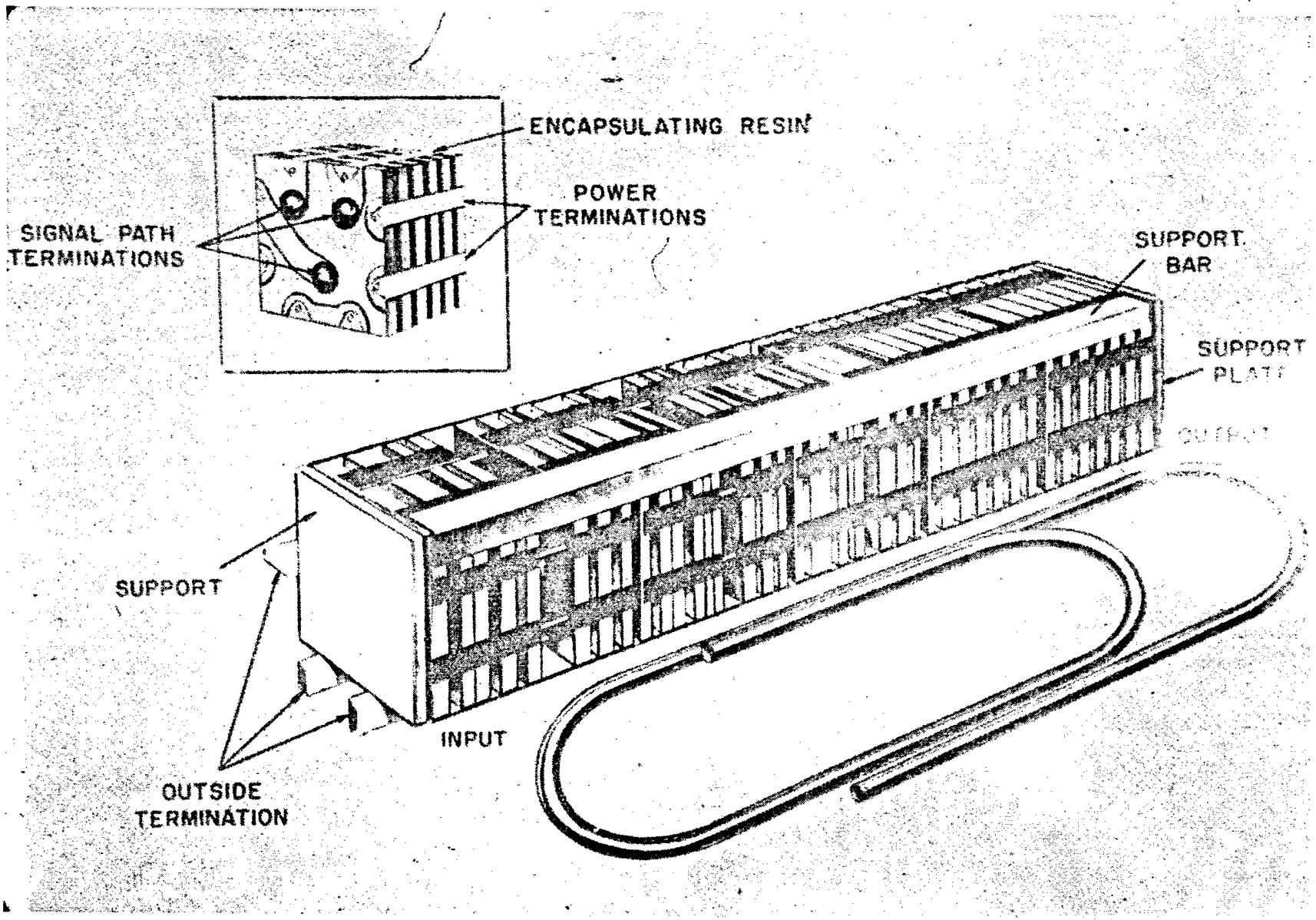
I have several slides which show some of these. The first slide is a miniaturized component as you see it along side of a paper clip. This isn't just one component - it has the entire circuit built inside of it and performs functions such as switching, oscillation and amplification.

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.

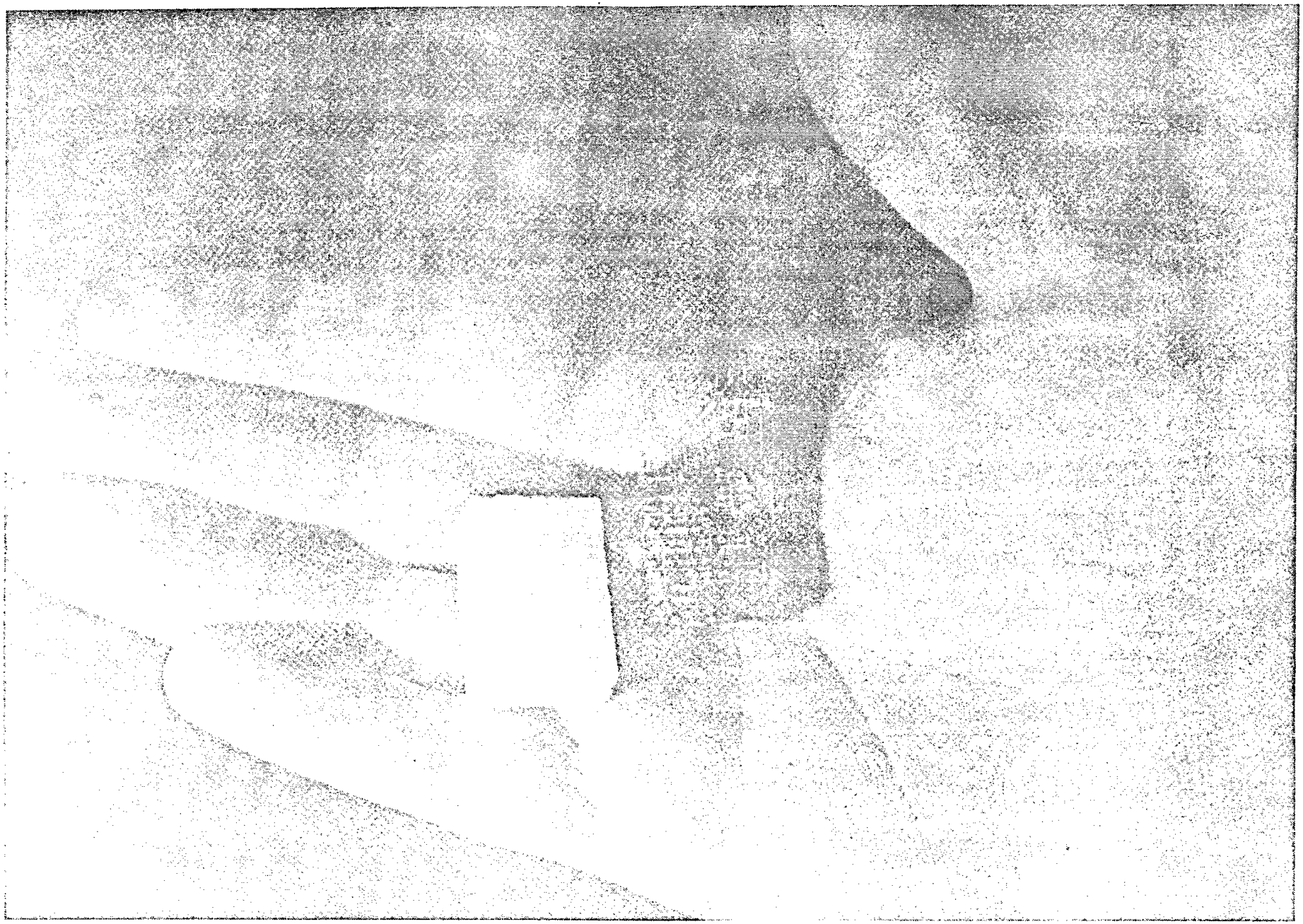
Slide 25

Micro modules



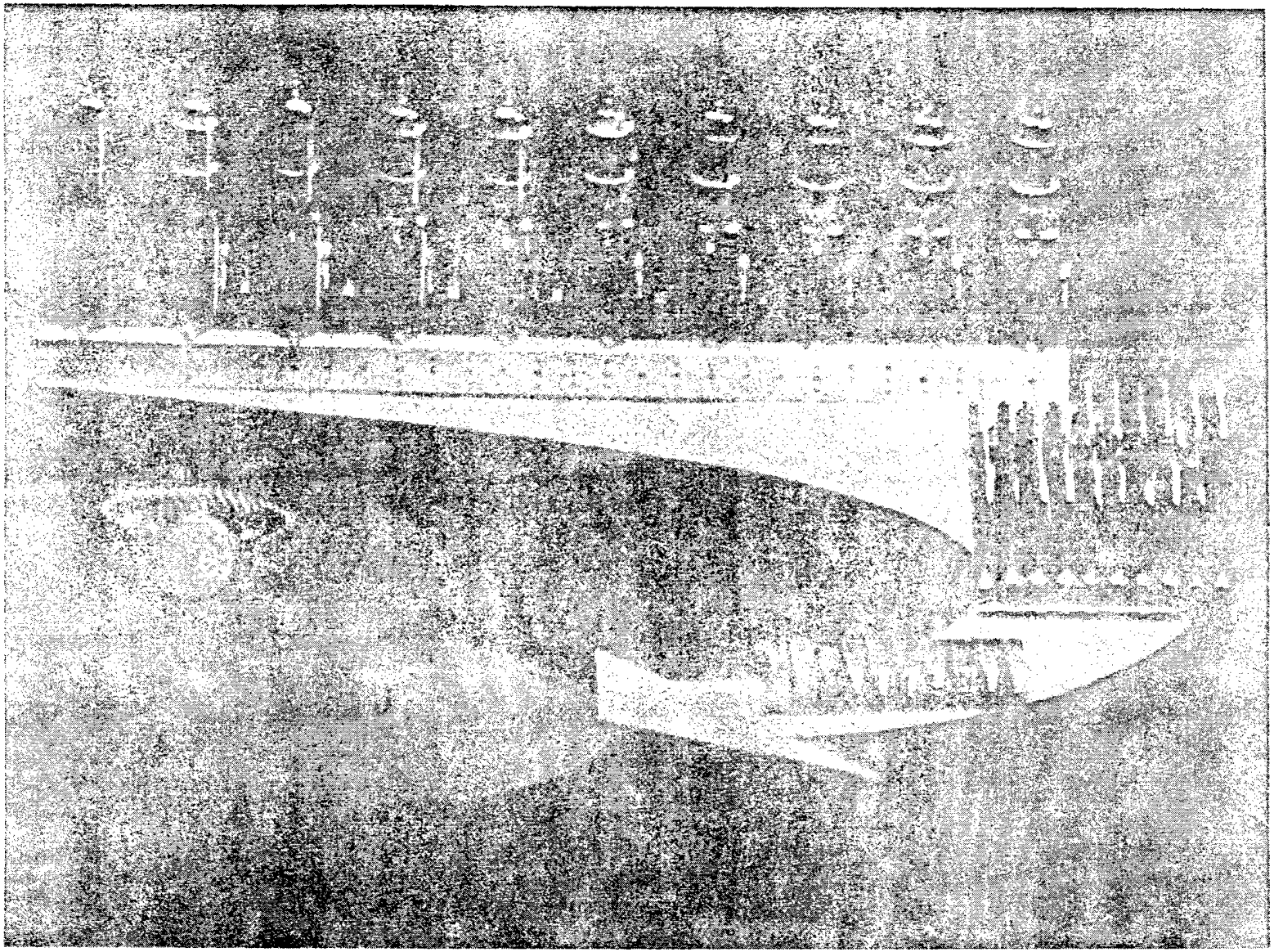
Slide 26

Micro module



Slide 27

Binary Computer

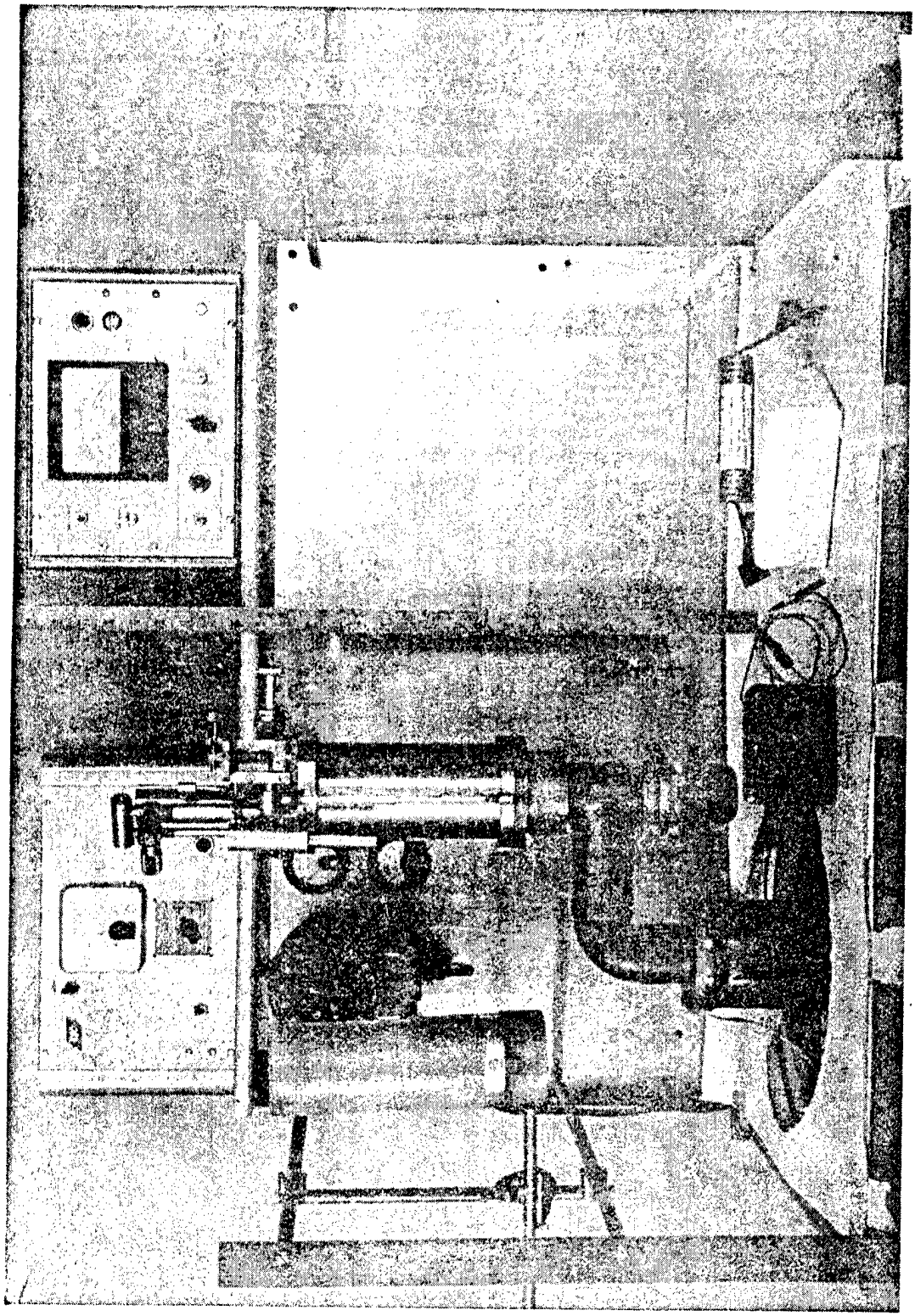


Slide 27

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

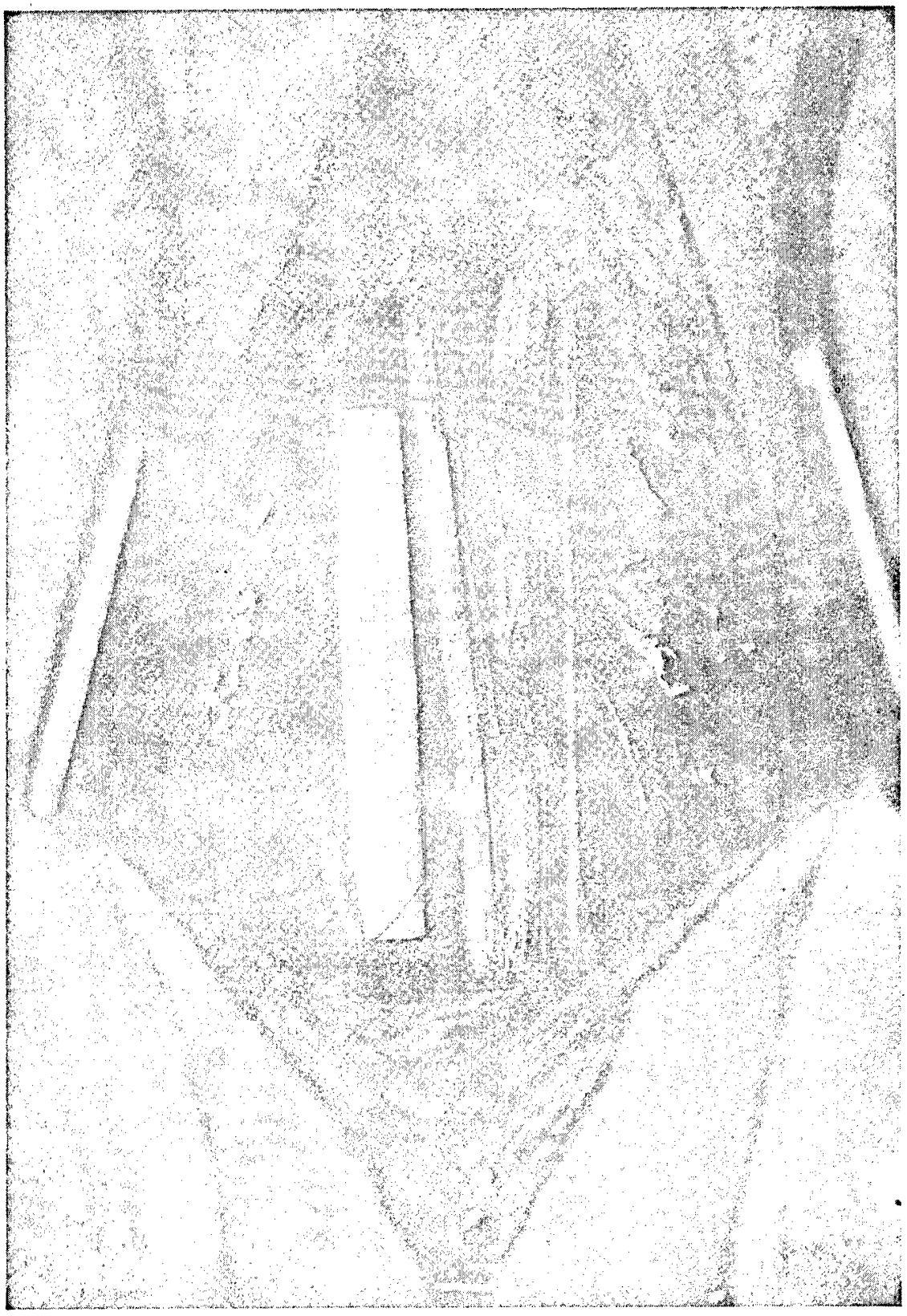
Slide 28

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.



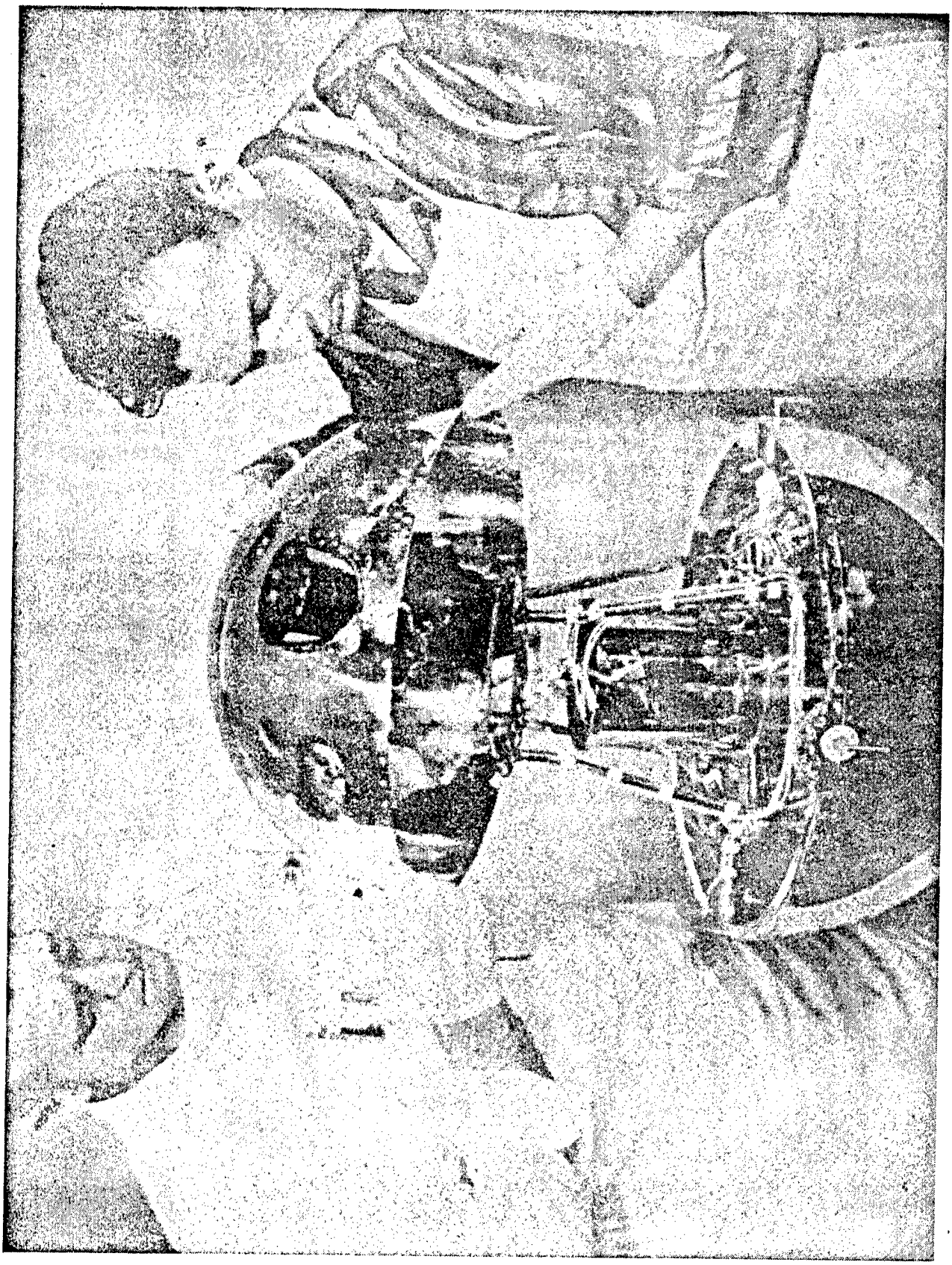
Slide 29
Nerve Repair

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.

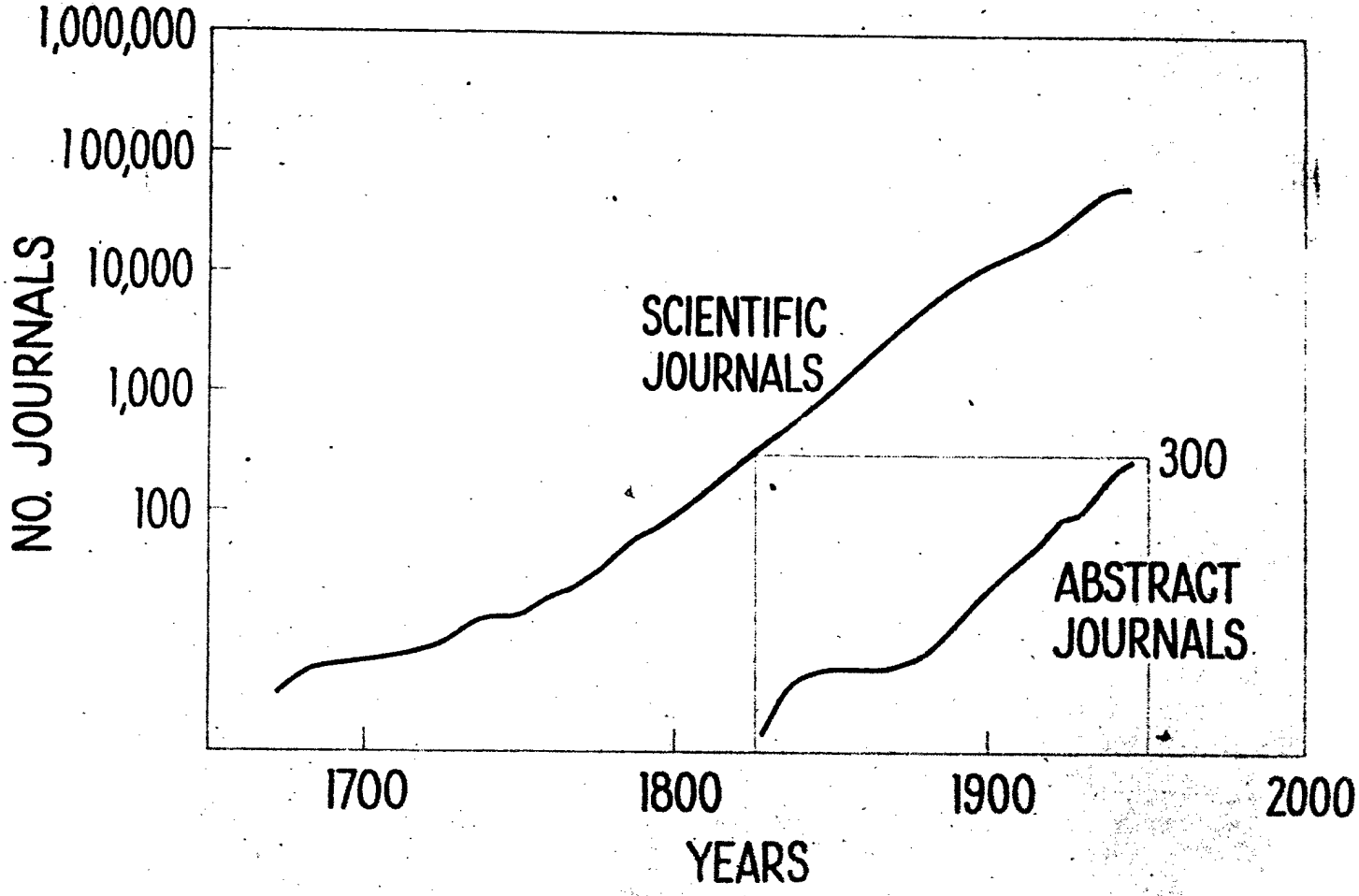
Just a brief summary; I think that one might gather, and I don't have to tell you people, because not only are you representatives of the Army but of industry, and you know that much of the research dollar that goes into defense ultimately finds useful outlets into civilian economy. The taxpayer certainly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

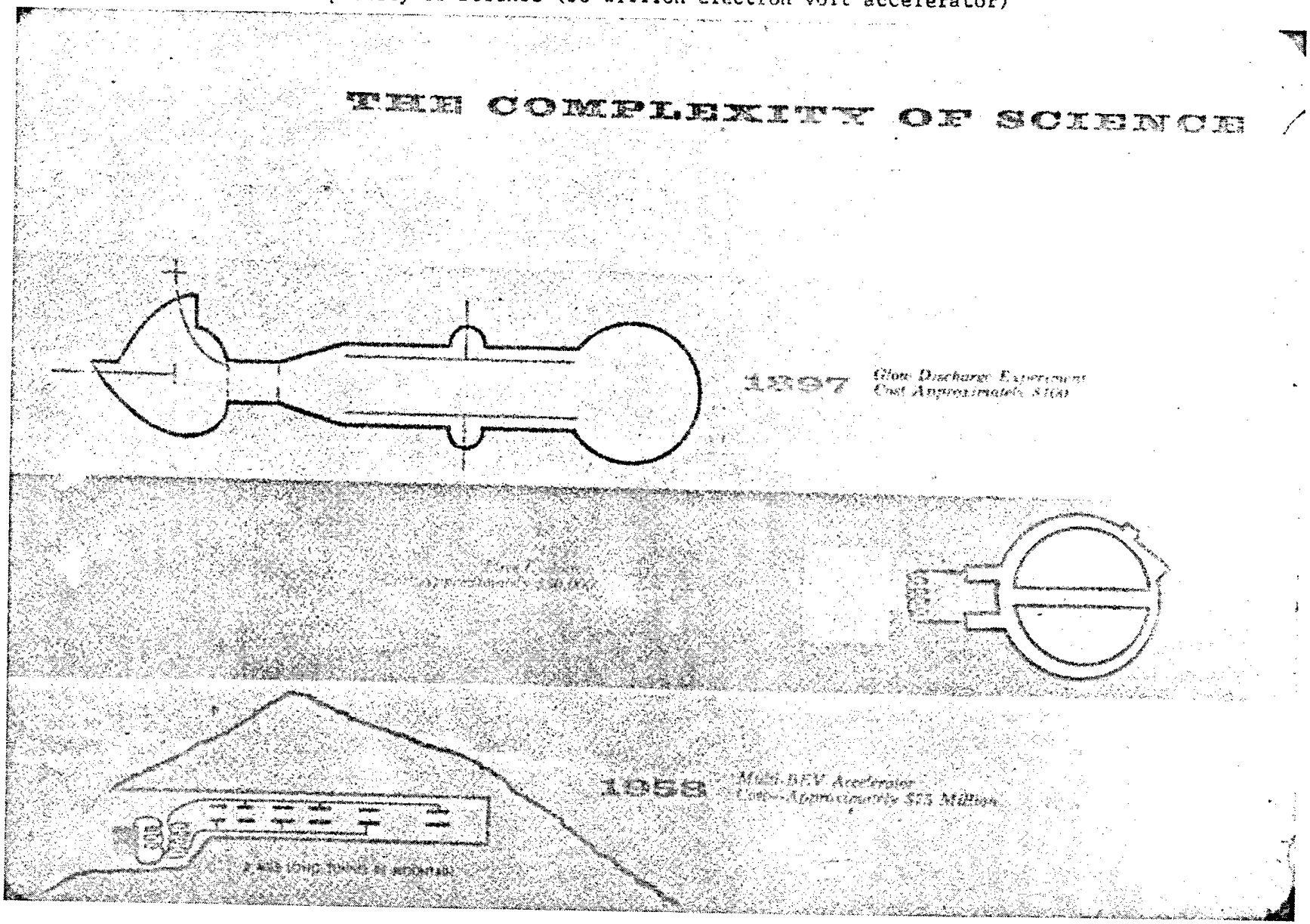


Slide 31

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.

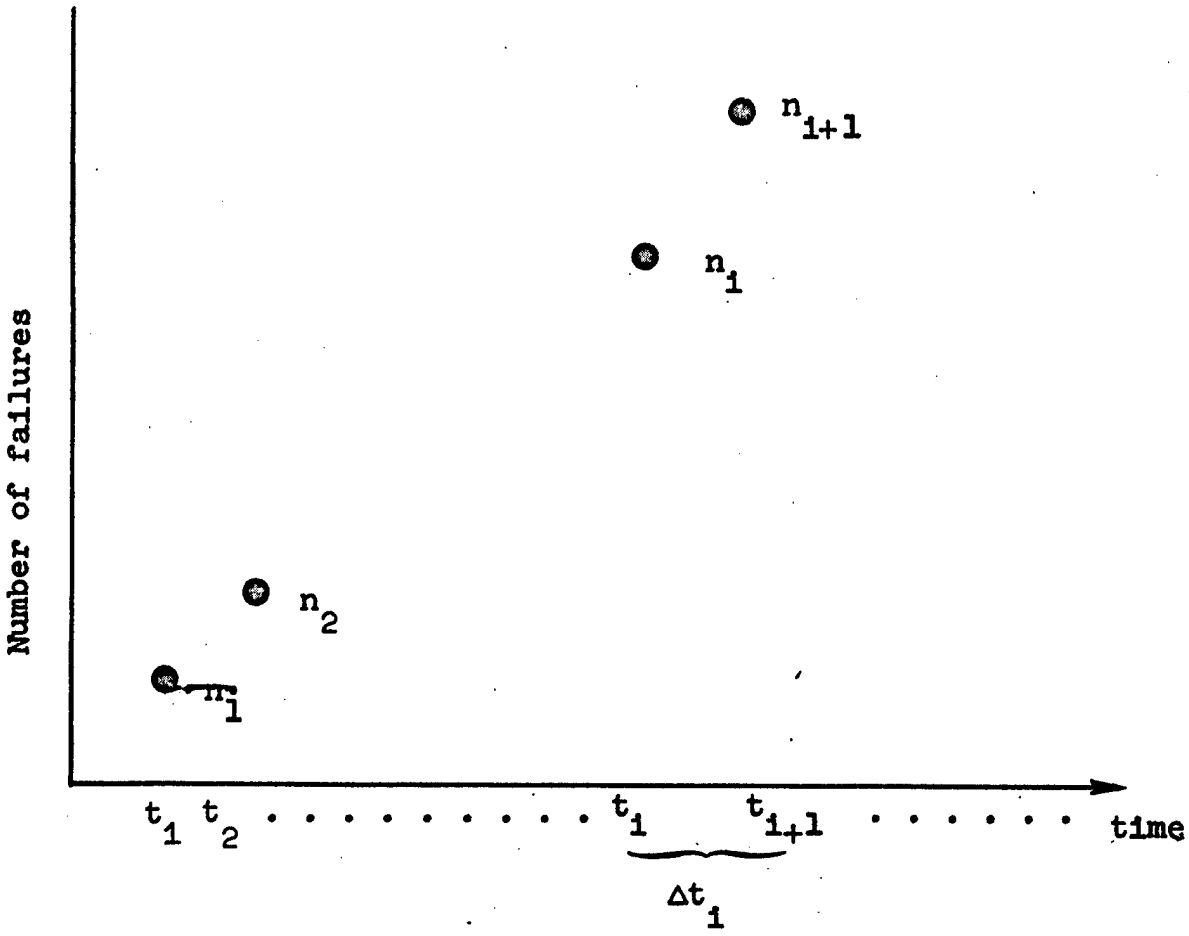
The approach I have outlined has been developing fruitfully, particularly during the last two or three years, in a number of other organizations concerned with complex systems. The White Sands Missile Range, the Rand Corporation, Chance-Vought Aircraft, Convair-Pomona, and the Autonetics Division of 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:

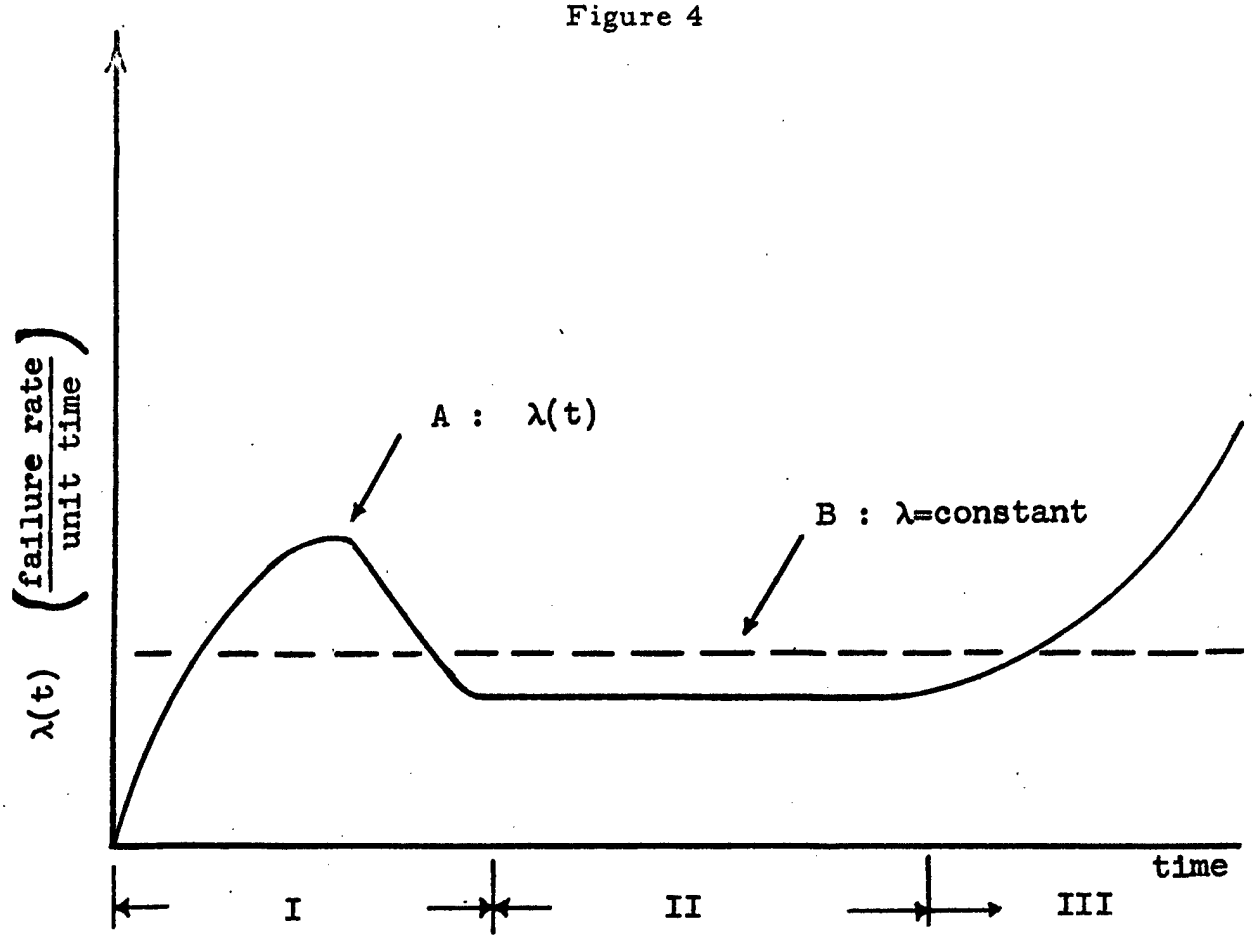
$$R(t) = R(o)e^{-\int_o^t \lambda(t)dt}$$

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

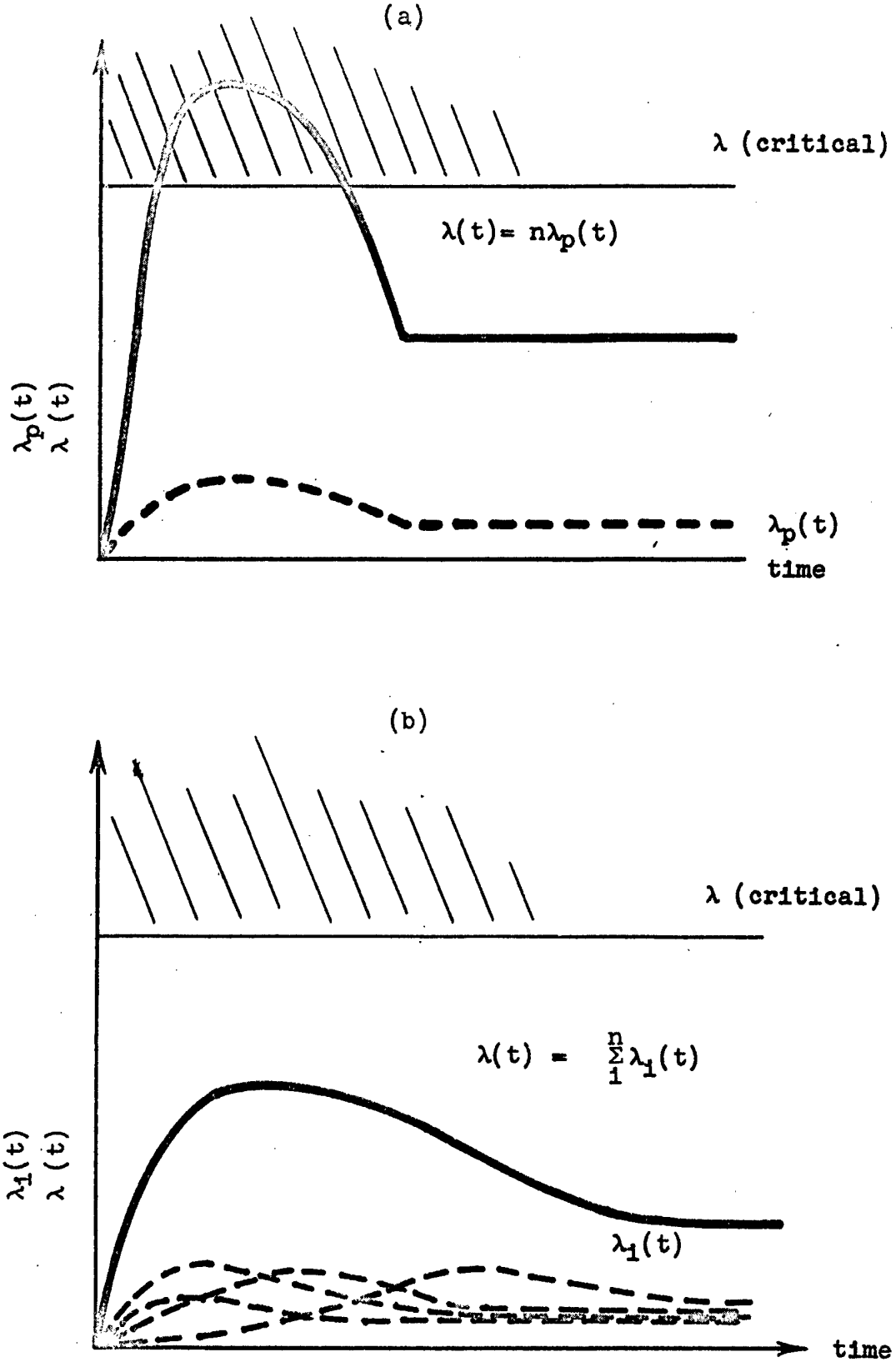
The straight line "B" is an average constant λ characteristic which, largely for the sake of simplicity in applications, is frequently assumed as applicable to most parts and components, for the purpose of taking, so to speak, a "first cut" at describing the corresponding reliability functions.

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).

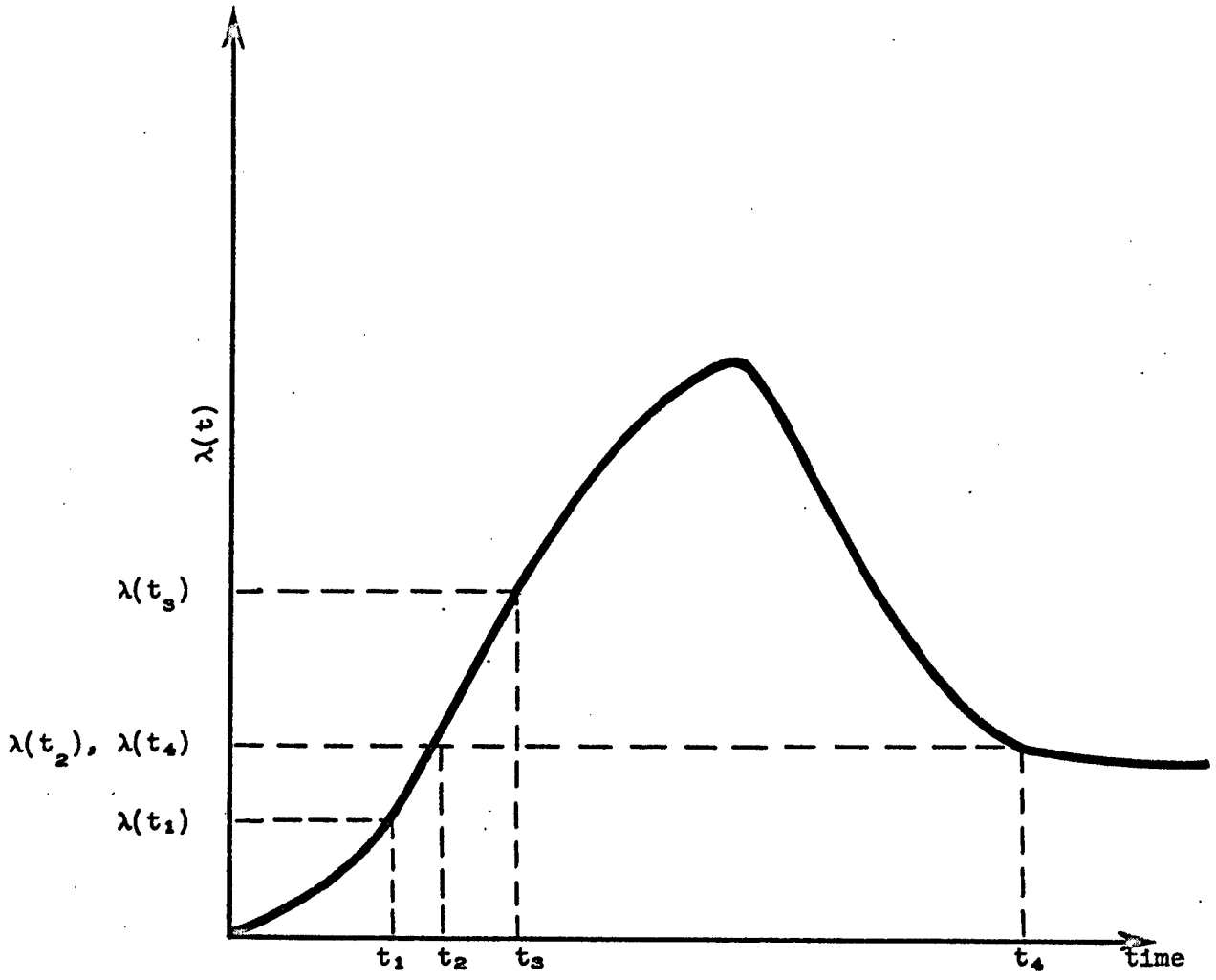
2. Secondly, if a number of different parts, with varying λ characteristics, are appropriately employed in a single component, it may be possible to design easily arrangements in critical circumstances that lead to a component λ characteristic which has desired form or a maximum desired failure rate 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:

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

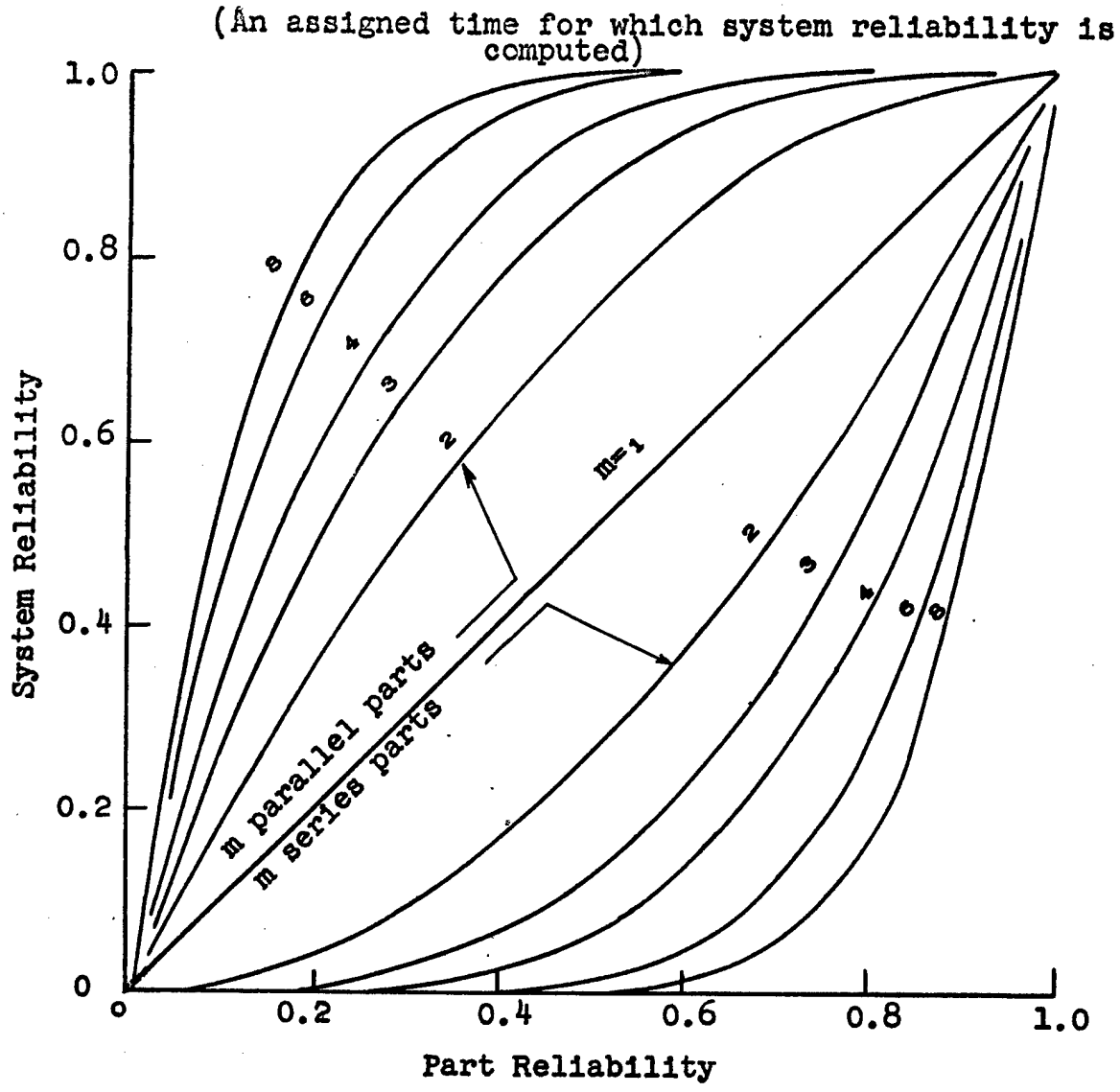
We are now familiar with the reliability function for a single part. If a component, for example, consists of n parts, which our analysis has shown to have independent failure probabilities, then the λ characteristic for it is simply the sum of the part characteristics and its reliability function is 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

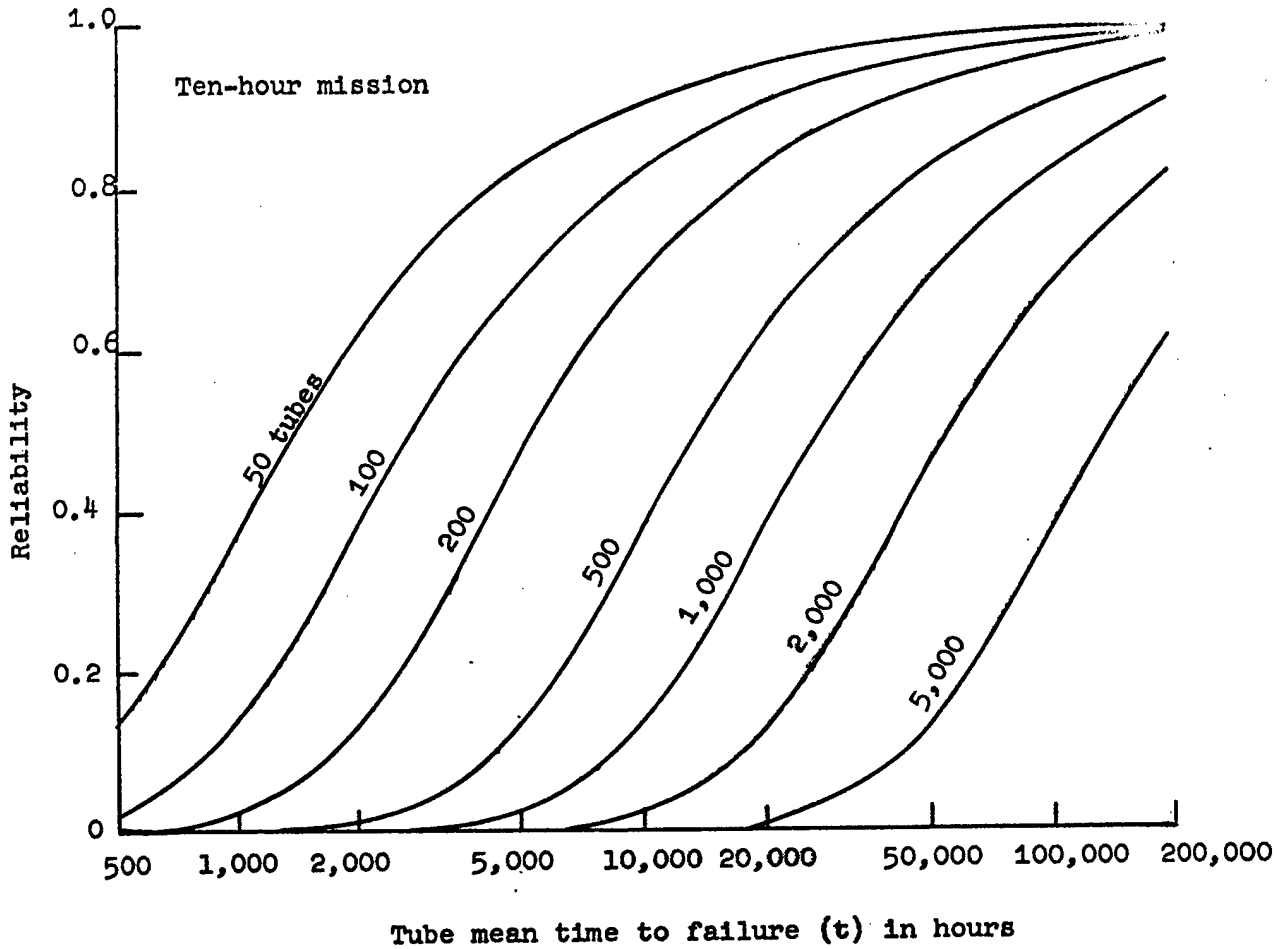


Reliability of a system of m identical parts

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."

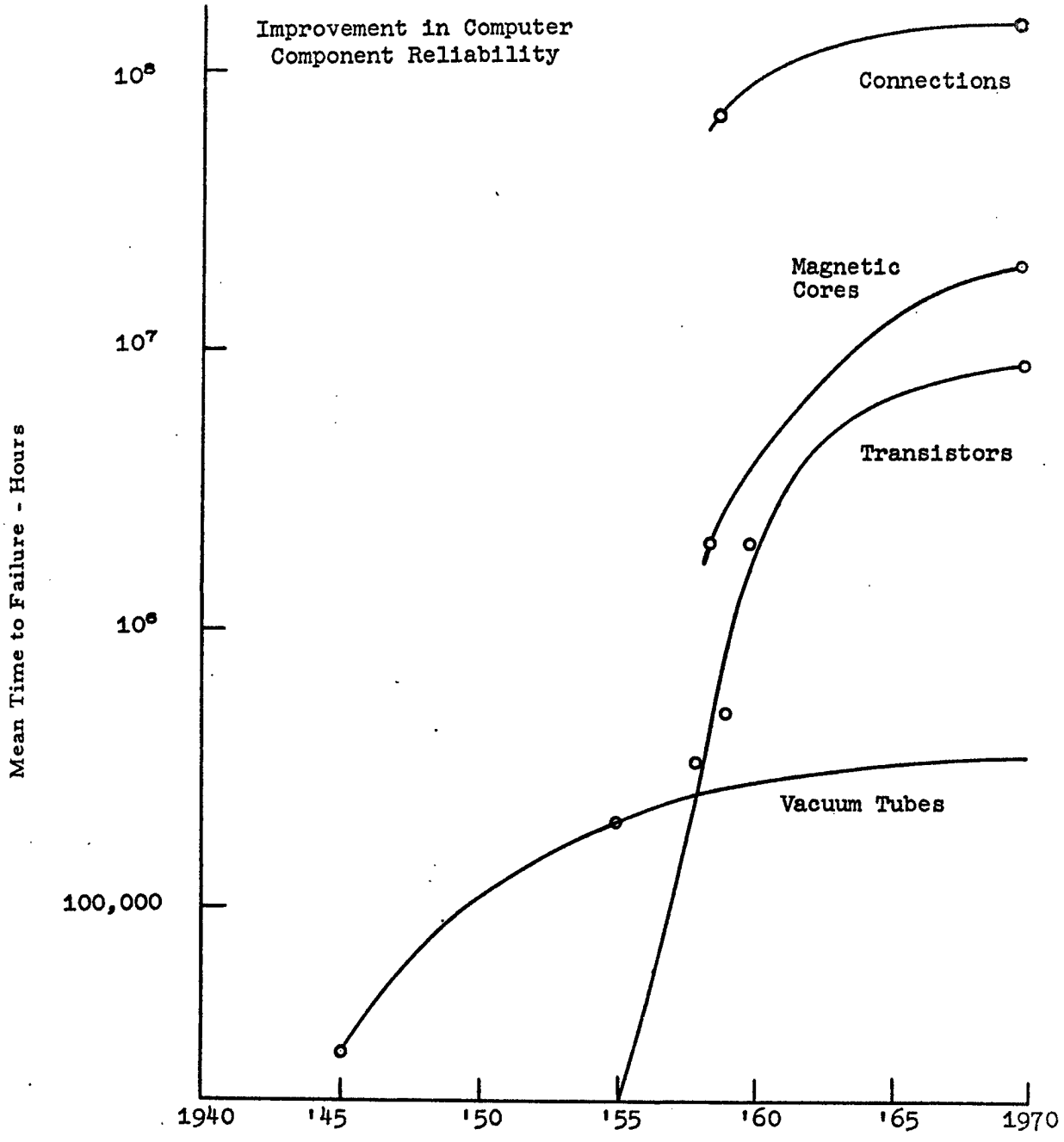
Figure 9



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

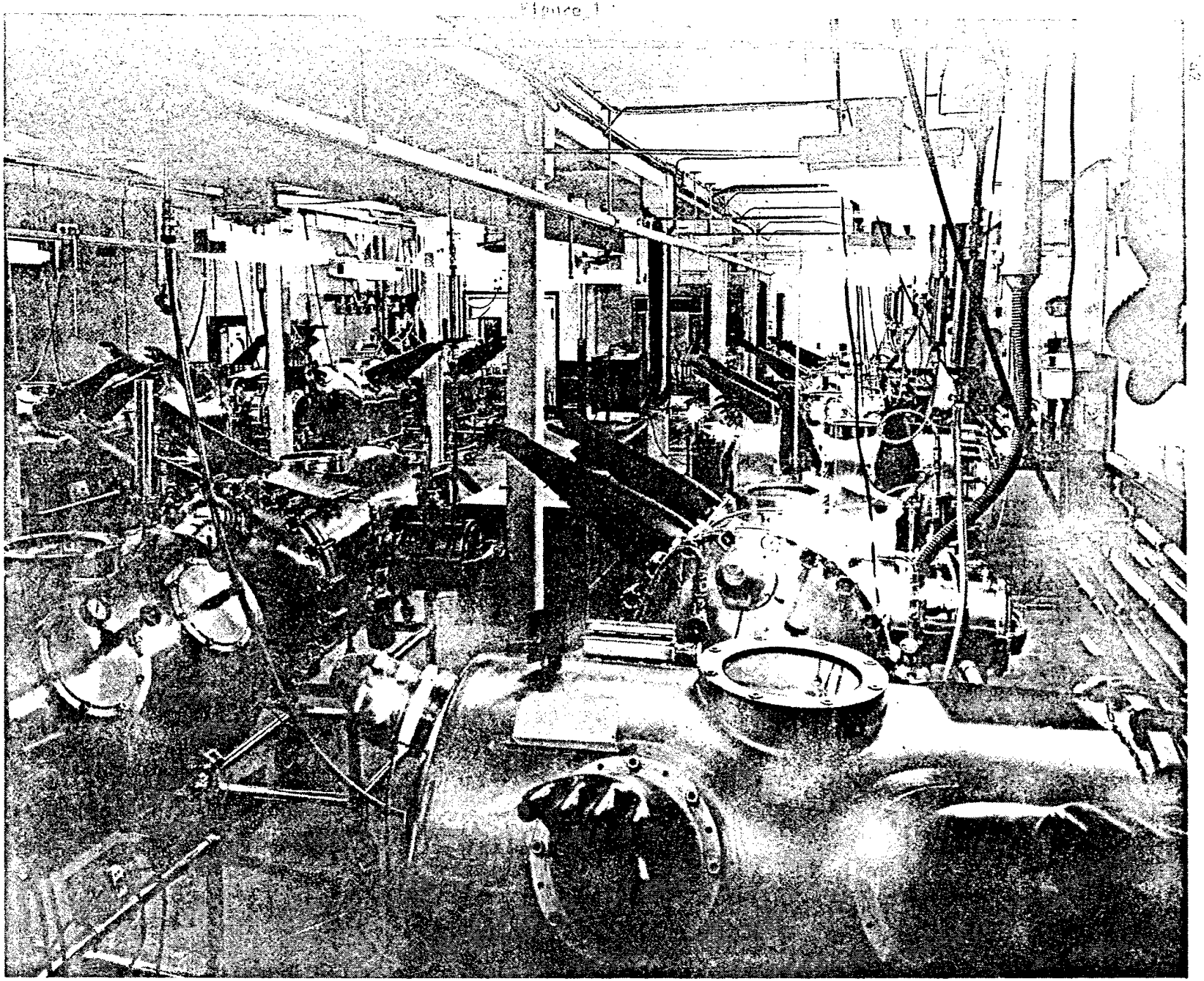


Figure 1

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

Figure 2

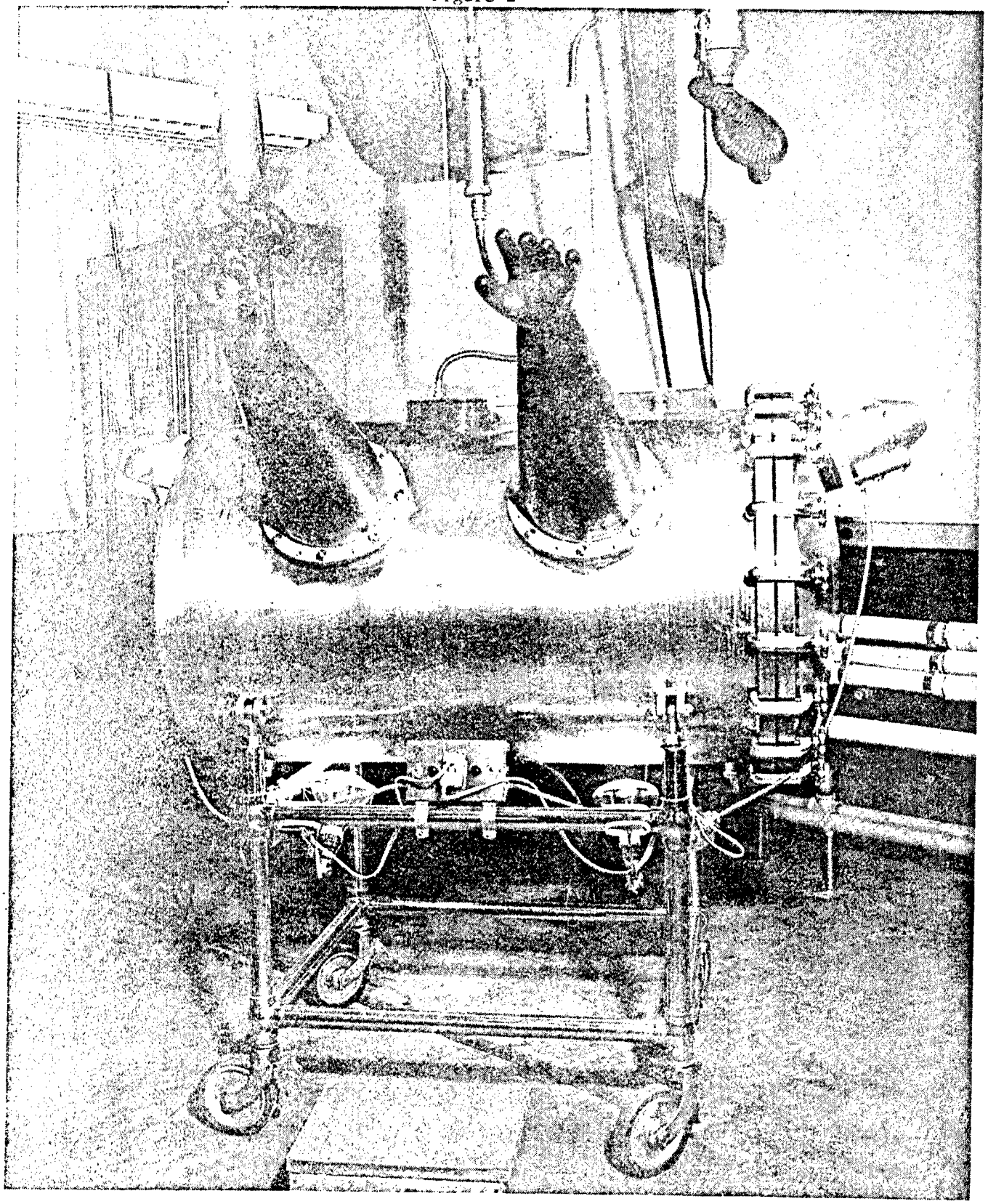


Figure 2

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

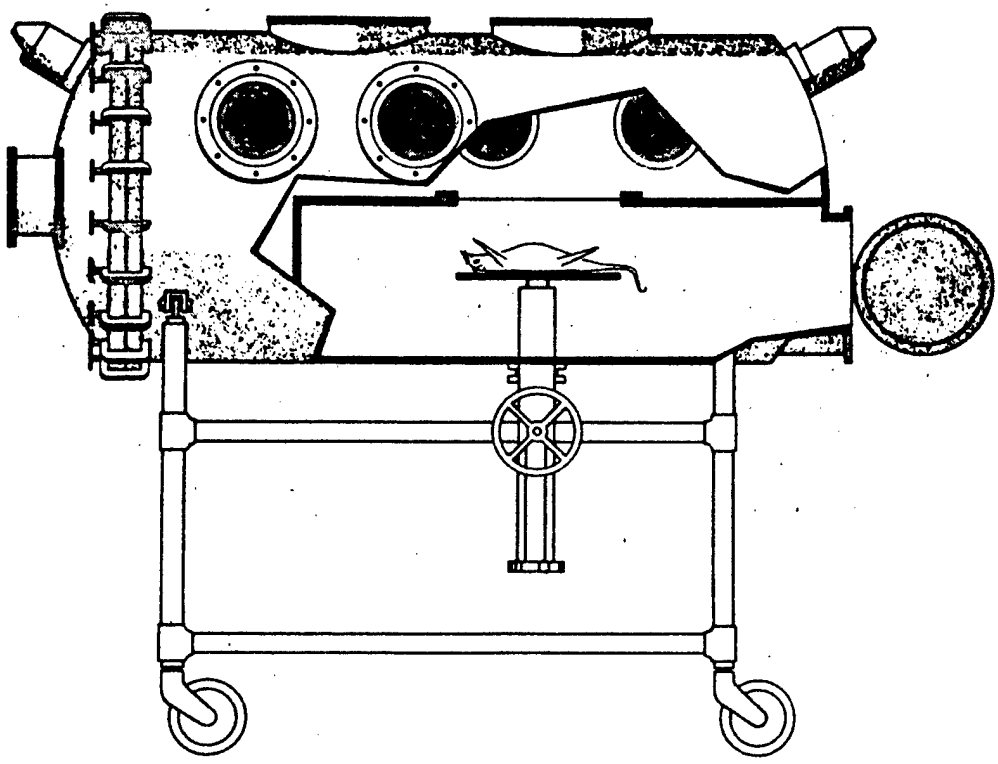


Figure 3

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.

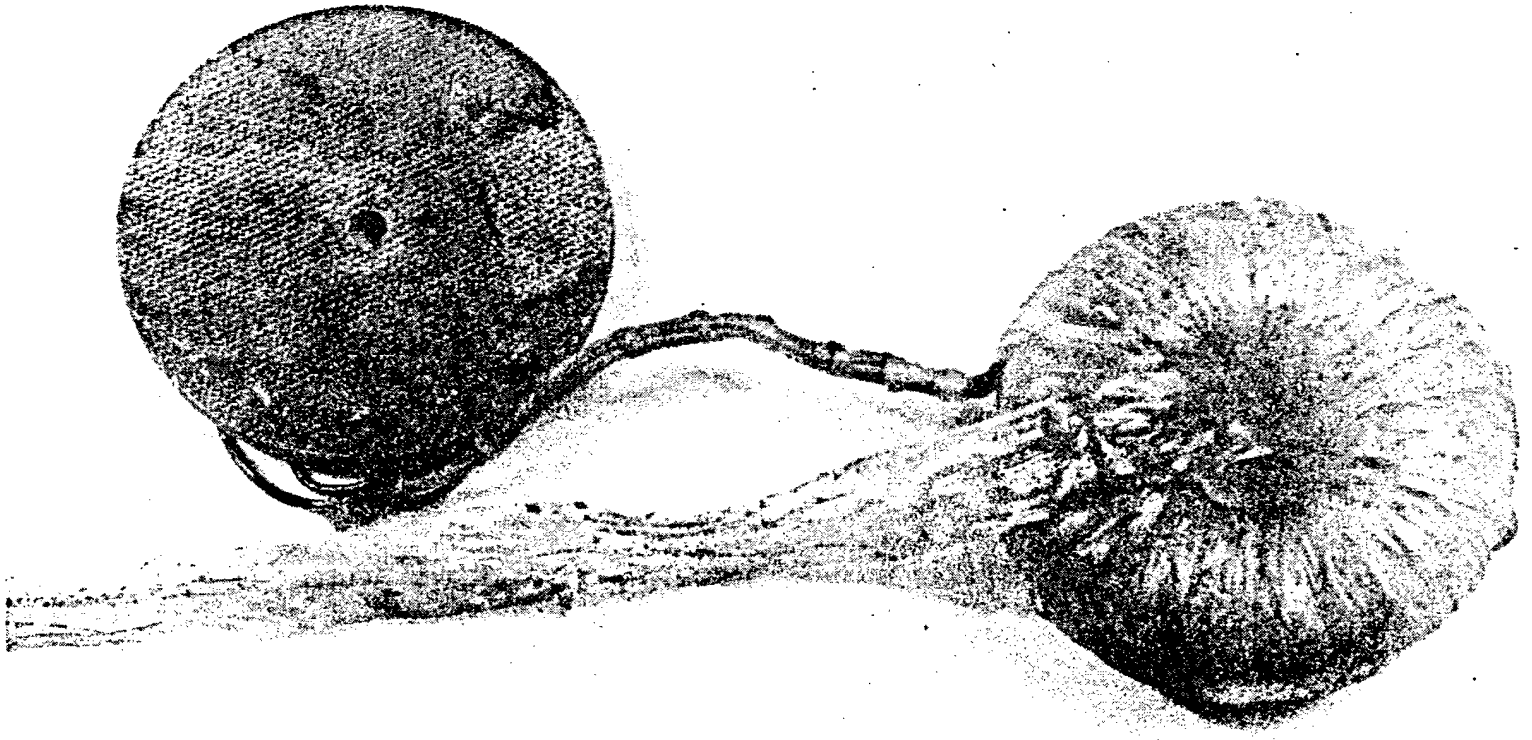
The requests of missiles manufacturers to our installation for information relative to their products' fungus resistance introduced a new phase of testing. Previous activities had been concerned with our mission for tank and tank-automotive vehicles and equipment. Usually, these materials were tested 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.

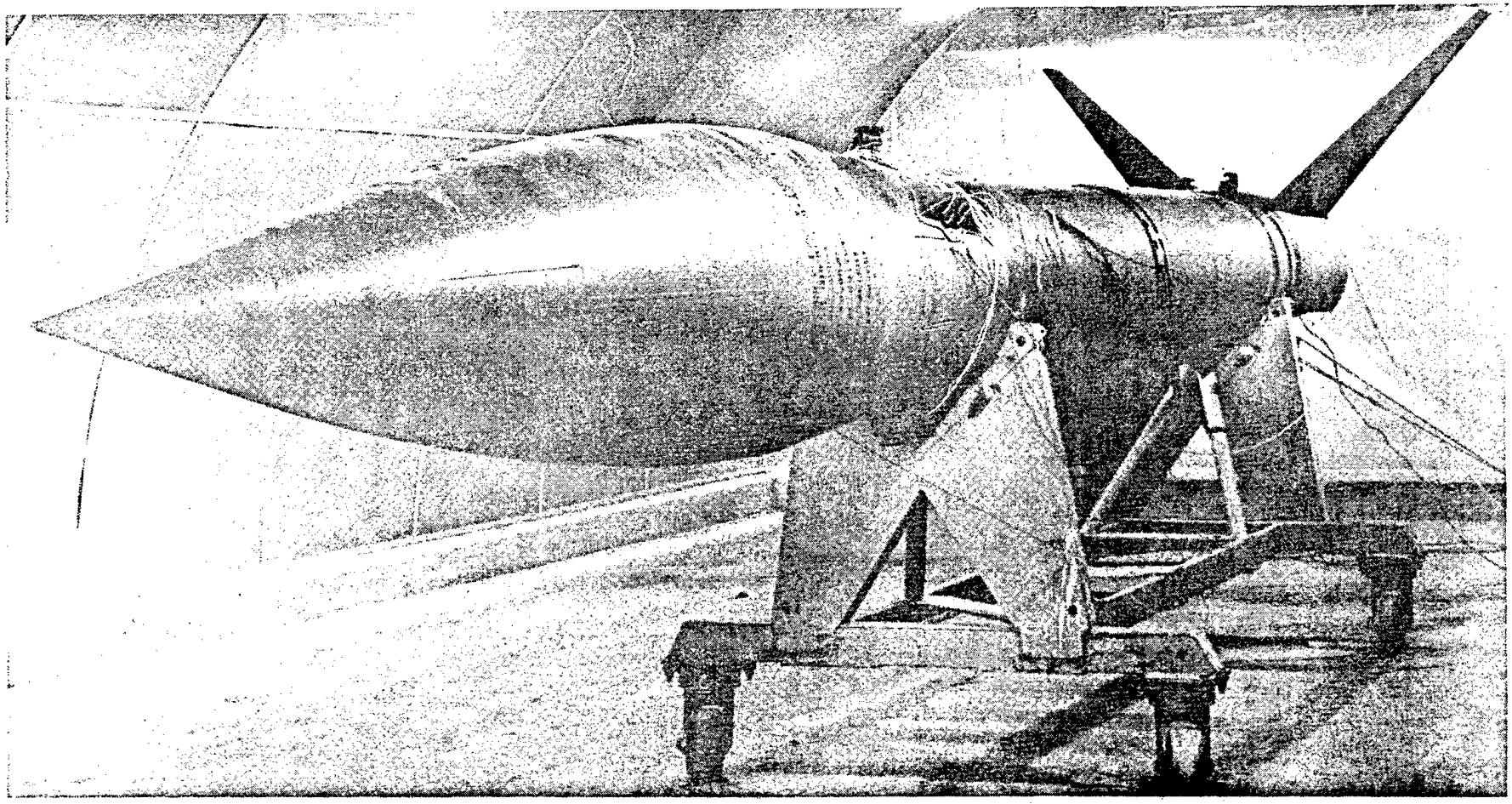
Literature surveys of the available specifications revealed no references specifically concerned with missiles, or the great variety of unique materials incorporated into these tactical weapons. Thus, our problem for the past few years was clearly indicated; the development of test parameters that would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

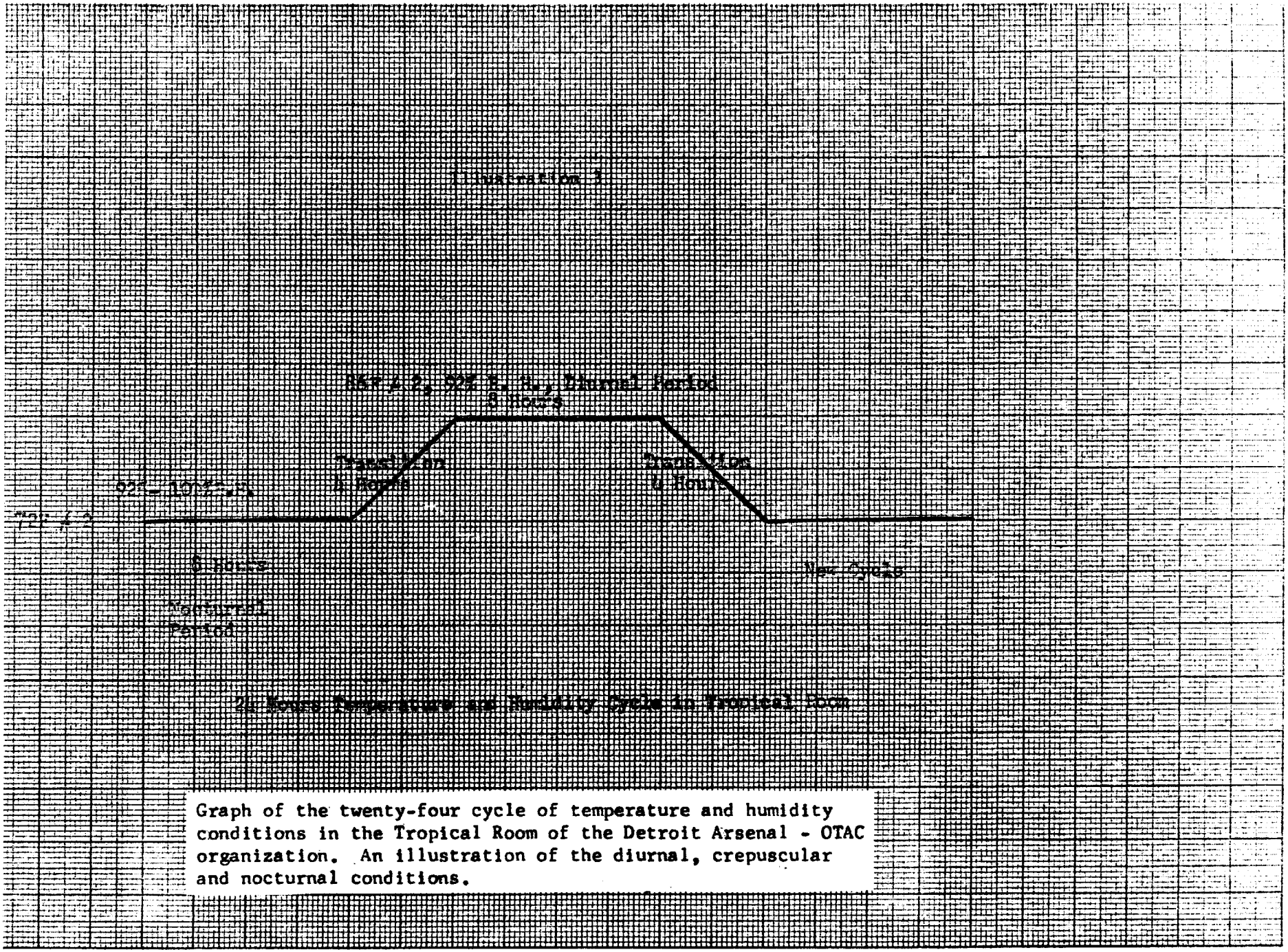
The conditions of temperature and humidity are original with the Detroit group and were determined on the basis of data available from the meteorological records of the rain forests of the world.

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).

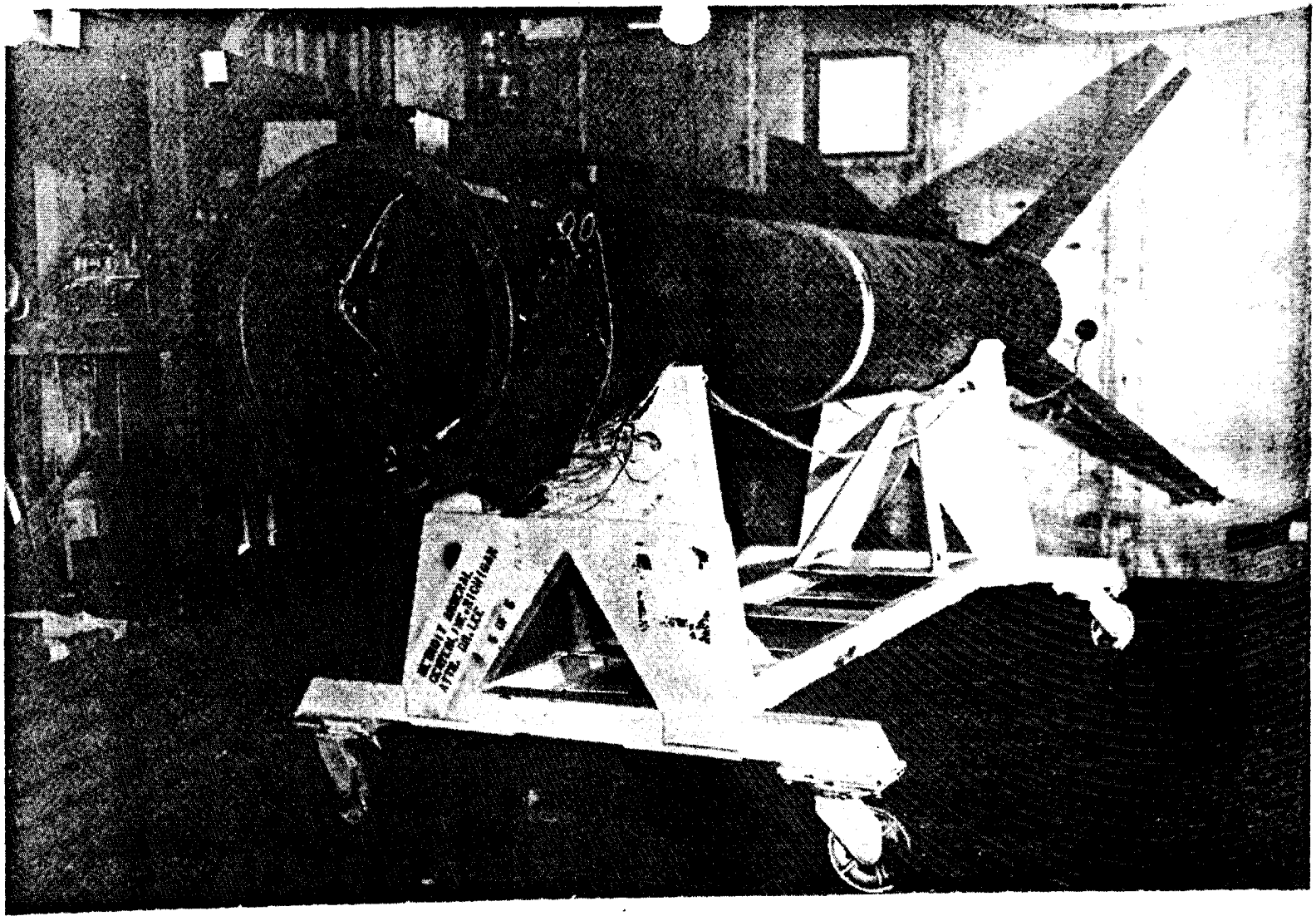
The use of the cycled atmosphere was evaluated by comparison and it was determined that with its use a greater number of species would be noted than using the room with constant temperatures and humidities.

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.



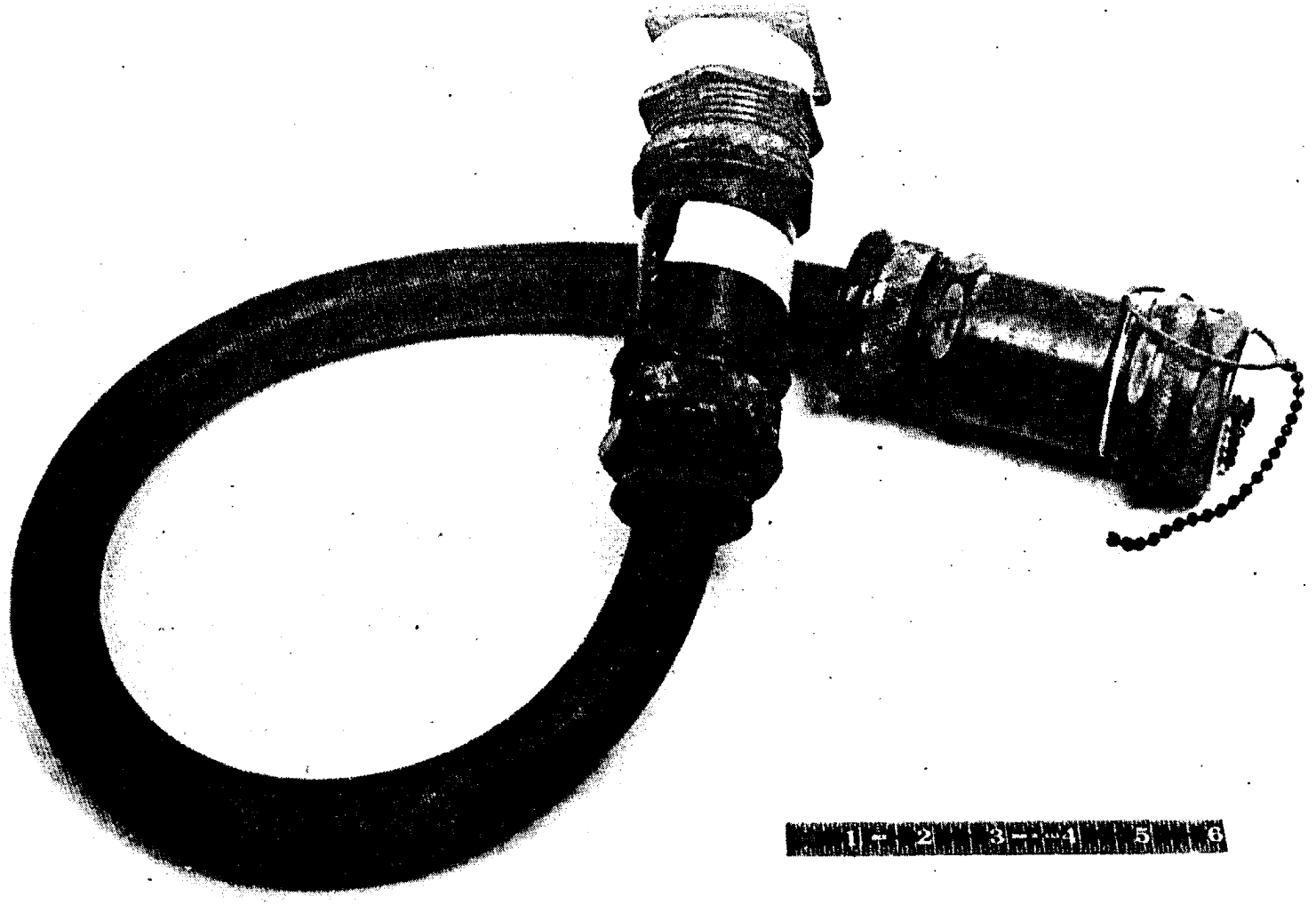
Graph of the twenty-four cycle of temperature and humidity conditions in the Tropical Room of the Detroit Arsenal - OTAC organization. An illustration of the diurnal, crepuscular and nocturnal conditions.

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.

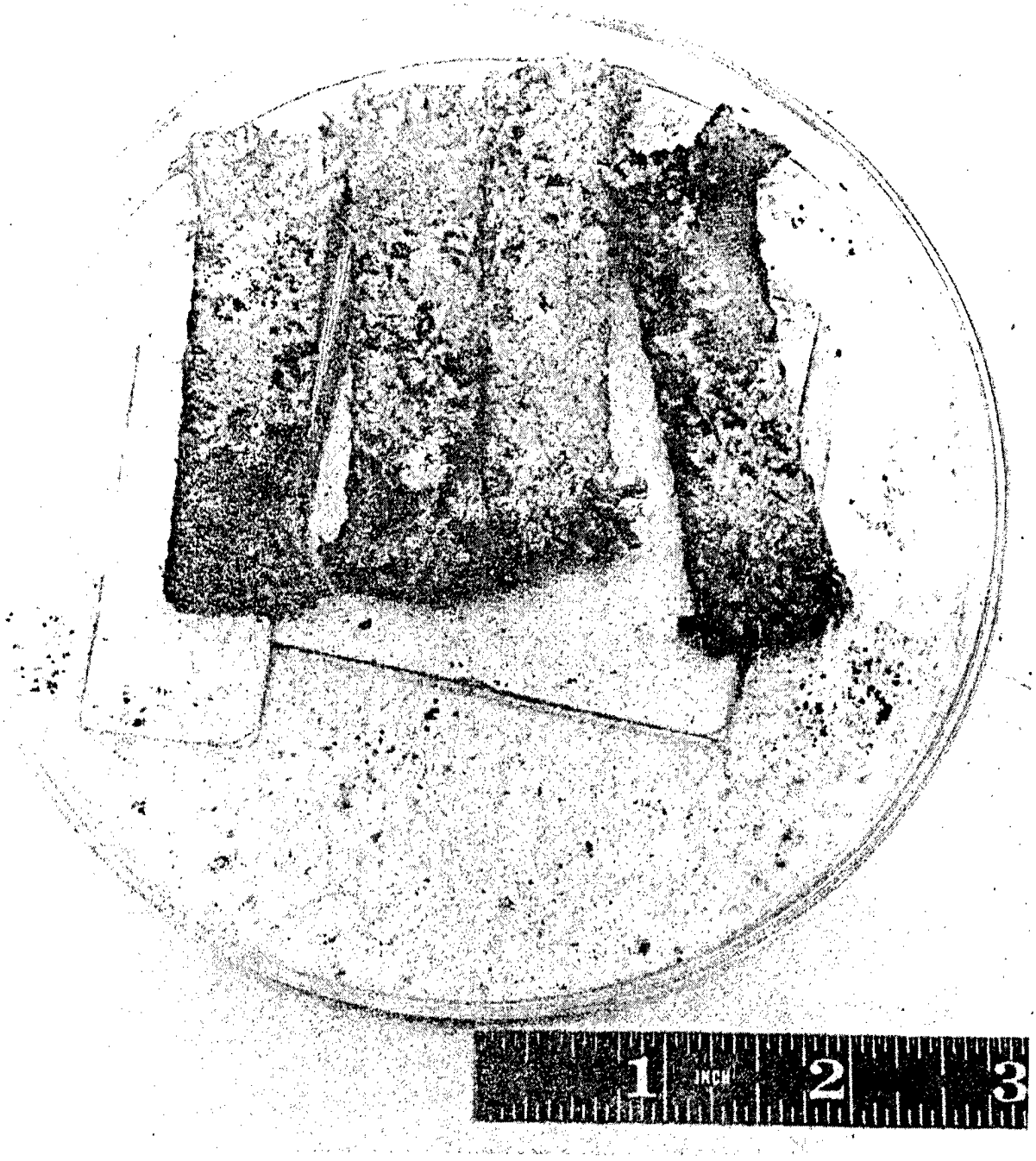
The fourth parameter developed for missiles testing was the choice of species of fungi. Again, in this matter, we relied on the data from past research and testing. However, since the majority of missiles items submitted for testing were new and unique, efforts had to be expended in securing information 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).

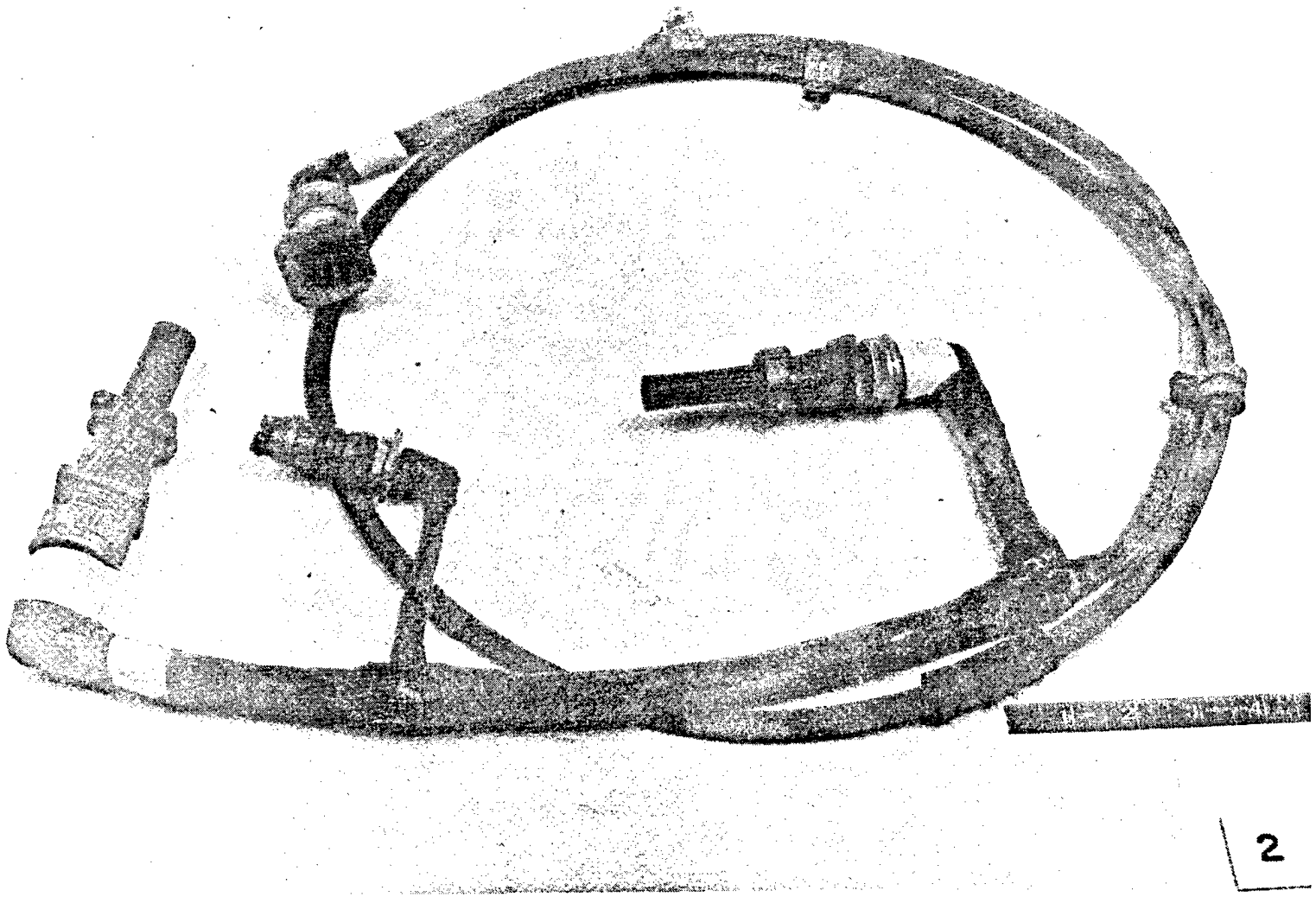
Support for the longer missile testing period relied on data secured on many dissimilar materials. In missiles work, it was noted that many of the items required a period for becoming conditioned to the atmosphere of the

Illustration 6



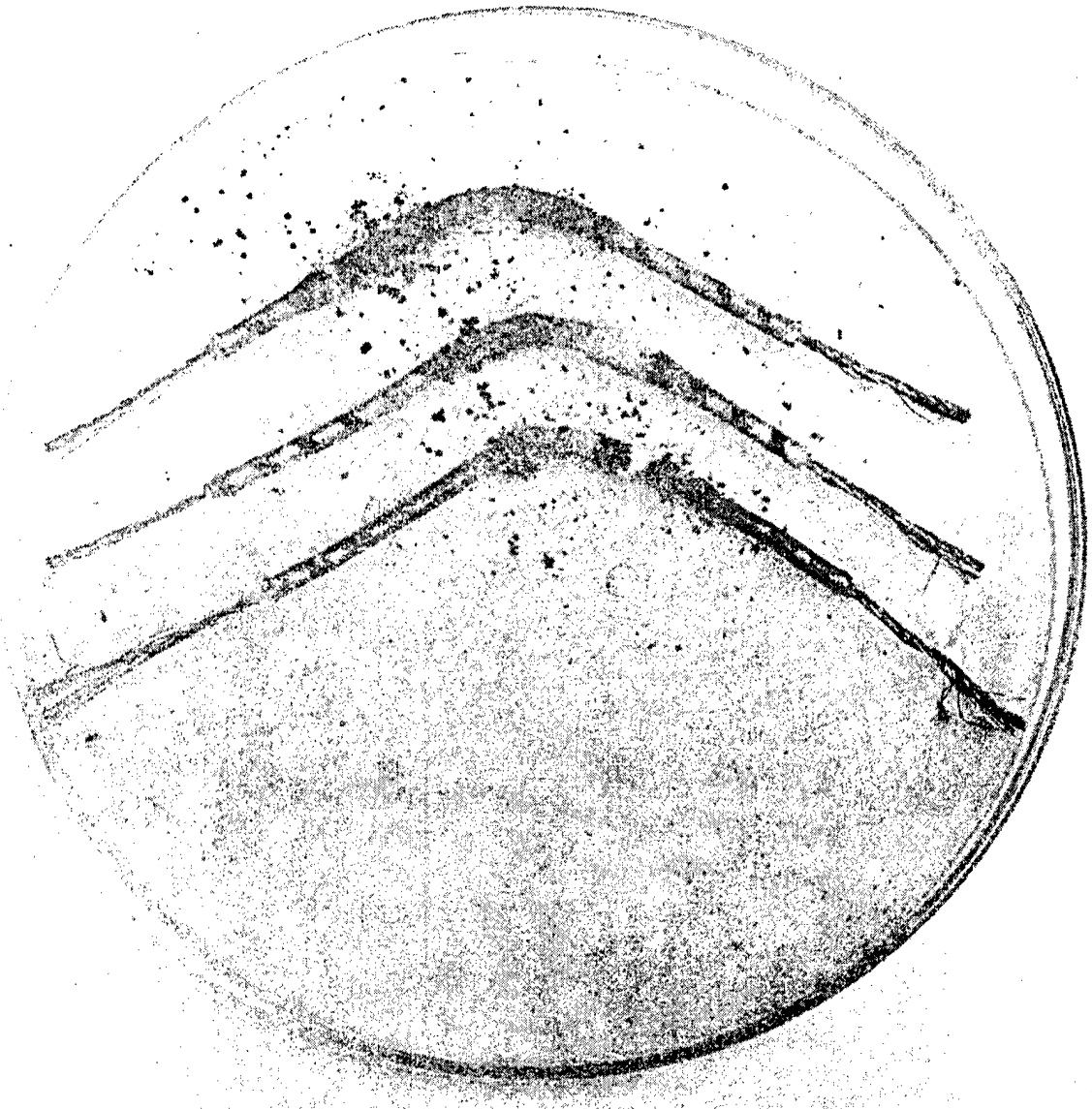
Asbestos samples adulterated with cotton. The cotton, being cellulose, supports a heavy growth of fungi while the funginert asbestos remains free of microbiological growth. Photograph taken at the conclusion of 90 days exposure in the Tropical Room at the Detroit Arsenal - OTAC.

Illustration 7

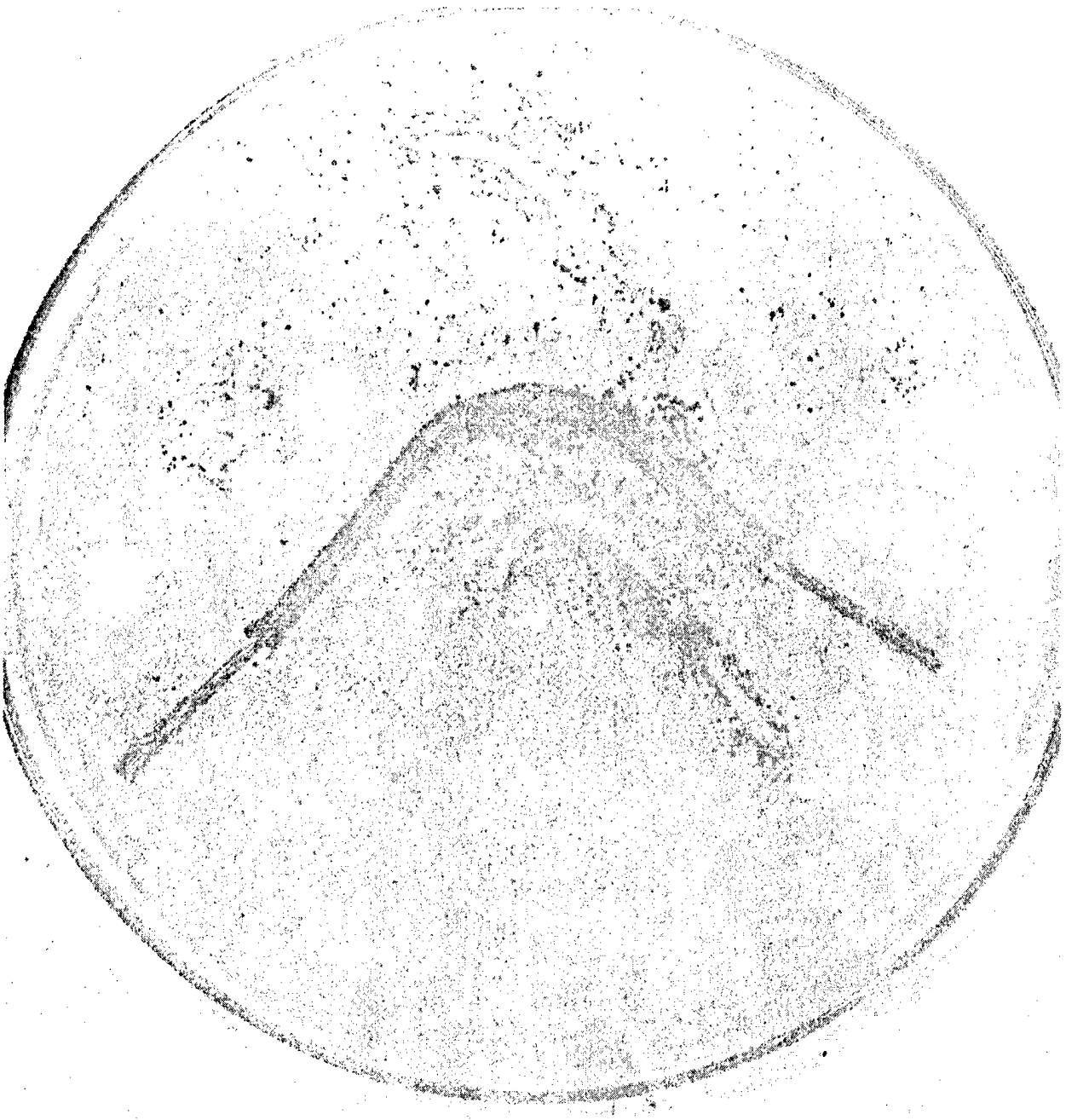


Cable harness following ninety days incubation with fungi in the Tropical Room of the Detroit Arsenal - OTAC. Cable insulation completely degraded and performance rating failed to meet the requirements for the component. Choice of inert materials is requisite.

Illustration 8

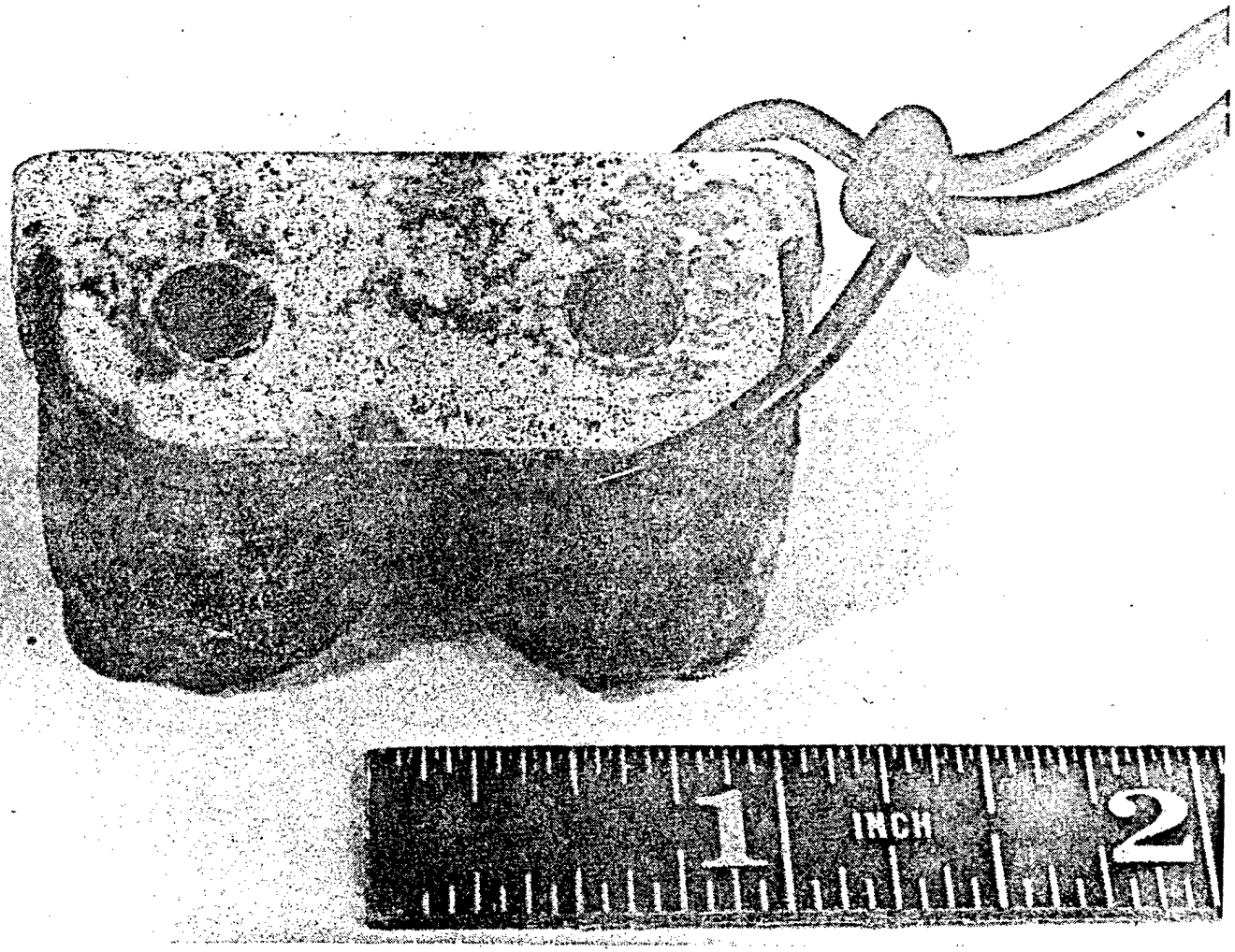


Cable leads following thirty days incubation in the Tropical Room of the Detroit Arsenal - OTAC organization. Both material resistance to fungus growth and performance ratings failed to meet minimum standards.



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

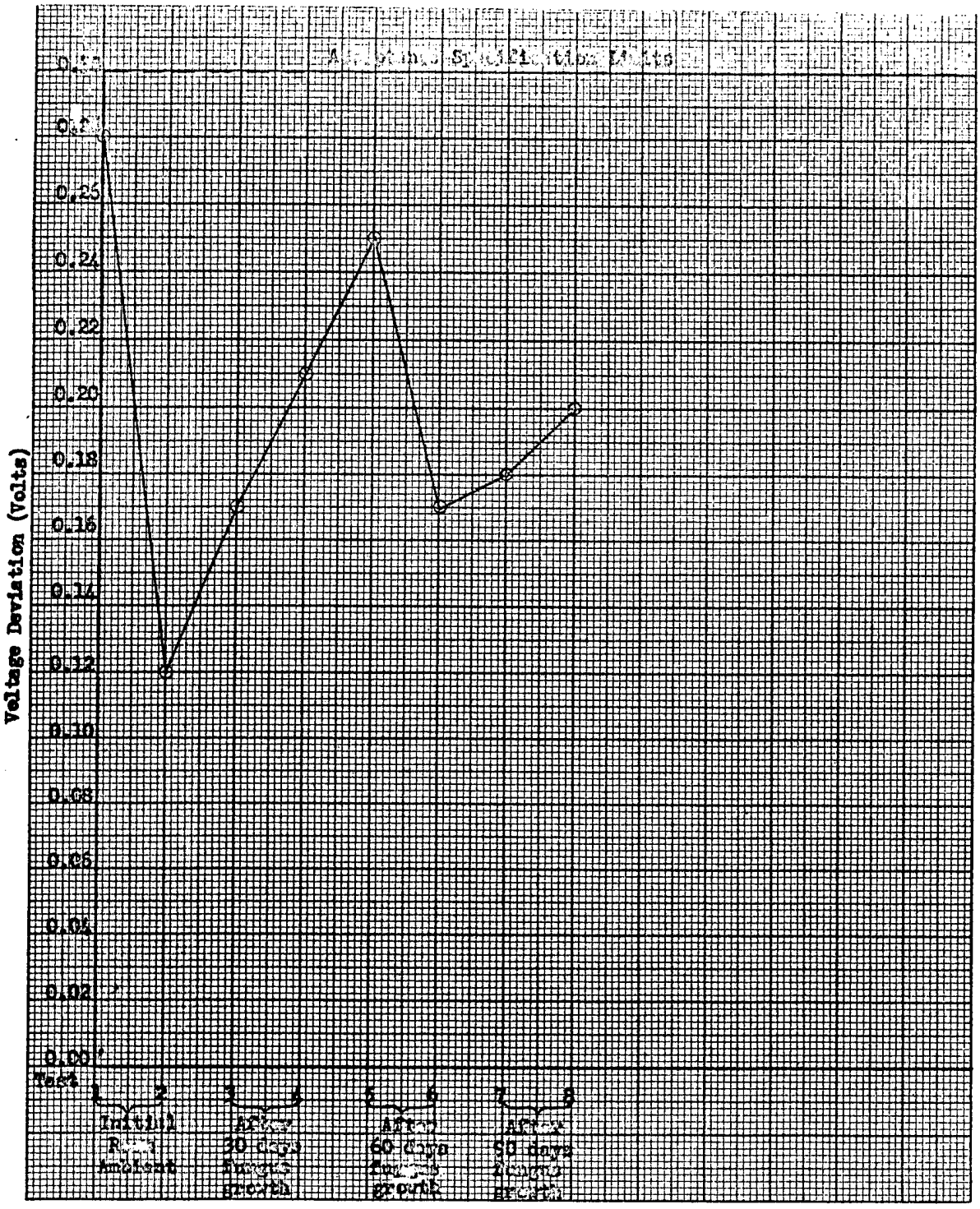


Figure 7 - Voltage deviation for constant-input, constant-load test (two-hour period)

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