

U. S. Army Research Office

Report No. 84-2

June, 1984

PROCEEDINGS OF THE TWENTY-NINTH CONFERENCE
ON THE DESIGN OF EXPERIMENTS

Sponsored by the Army Mathematics Steering Committee

HOST

Uniformed Services University of Health Sciences

Bethesda, Maryland

19-20 October 1983

Approved for public release; distribution unlimited.
The findings in this report are not to be construed
as an official Department of the Army position, un-
less so designated by other authorized documents.

U. S. Army Research Office
P. O. Box 12211
Research Triangle Park, North Carolina

FOREWORD

The Twenty-Ninth Conference on the Design of Experiments in Army Research, Development and Testing was held October 19-21, 1983, at the Uniform Services University of Health Sciences (USUHS), Bethesda, Maryland. This was the second Army-wide conference to be held at this university. The first one, called the Twenty-Eighth Conference of Army Mathematicians, was held June 28-30, 1982. As a result of this June meeting, Dr. David Cruess, a faculty member of USUHS, offered the facilities of his university for the Twenty-Ninth Conference on the Design of Experiments. The members of the Army Mathematics Steering Committee (AMSC), sponsors of these conferences, were pleased to receive this invitation. They would like to take this occasion to thank Professor Cruess for serving as Local Chairperson and for his excellent handling of the many problems associated with a meeting of this size. A brief history of USUHS appeared in a booklet issued to the attendees of this conference. This interesting and informative booklet is reproduced at the end of this Foreword.

Two days before the start of the Design Conference, a tutorial entitled, "Sequential Methods in Statistics," was held. Its speaker was Professor Michael Woodroffe of the University of Michigan at Ann Arbor, Michigan. The main purpose of this seminar was to develop, in Army scientists, an appreciation for and the necessary skills needed to handle some of the statistical methods for analyzing experimental data.

Members of the Program Committee for this conference were pleased to obtain the services of the following invited speakers to talk on topics of current interest to Army personnel:

<u>Speaker and Affiliation</u>	<u>Title of Address</u>
Dr. Marvin A. Schneiderman National Cancer Institute	EPIDEMIOLOGY AND RISK ASSESSMENT: COURTS, CLOCKS AND CONFUSION
Dr. William Sacco Washington Hospital Center	INJURY SEVERITY SCORES AND APPLICATIONS TO MILITARY TRIAGE
Professor Jerome Friedman Stanford Linear Accelerator Center	INTERACTIVE COMPUTER DATA ANALYSIS
Dr. Charles Brown National Cancer Institute	HIGH TO LOW DOSE EXTRAPOLATION OF EXPERIMENTAL ANIMAL CARCINOGENESIS STUDIES

A broad overview of the many research areas presented in the contributed papers can be ascertained from the titles of the various sessions:

Special Session:	Sequential Testing
Technical Session 1:	Statistical Theory
Technical Session 2:	Analysis of Longitudinal Data
Technical Session 3:	Simulation Techniques and Applications
Technical Session 4:	Test and Evaluation Techniques
Technical Session 5:	Application in Experimental Design

In addition to the above mentioned sessions, there was a Clinical Session which offered an opportunity to each of three Army scientists to present unsolved statistical problems and receive suggestions and constructive comments from the experts.

Professor Herbert A. David of the Department of Statistics, Iowa State University, was the recipient of the third Wilks Award for contributions to Statistical Methodologies. He received this award at the banquet held at the Officer's Club, Naval Medical Center, on October 19, 1983. This honor was bestowed on Dr. David for his many significant contributions to various fields of statistics, in particular to the areas of order statistics and competing risks, and also for his contributions to the Army. He has assisted many Army scientists with their statistical problems, served as invited speaker at two Design conferences, and provided theoretical details for the solution of a fuzing problem for the Ballistic Research Laboratory.

The AMSC has requested that these Proceedings be published and distributed Army-wide so that the information it contains could assist Army scientists with some of their statistical problems. Committee members would like to thank the Program Committee for all it did in putting together this scientific conference.

Program Committee

Carl Bates
Larry Crow
David Cruess
Walter Foster
Bernard Harris
Robert Launer

Richard Moore
James Schlesselman
Douglas Tang
Malcolm Taylor
Jerry Thomas
Langhorne Withers

School of Medicine

ADMINISTRATION

Jay P. Sanford, M.D.
Dean, School of Medicine

Leonard W. Johnson, Jr., M.D.
COL, USAF, MC
Associate Dean

Peter J. Stavish, M.B.A., LTC, MSC, USA (Ret)
Director of Admissions
Registrar

Joan F. Crotty, M.S.W.
Assistant Director of Admissions

INQUIRIES

For more information about the Uniformed Services University write to:

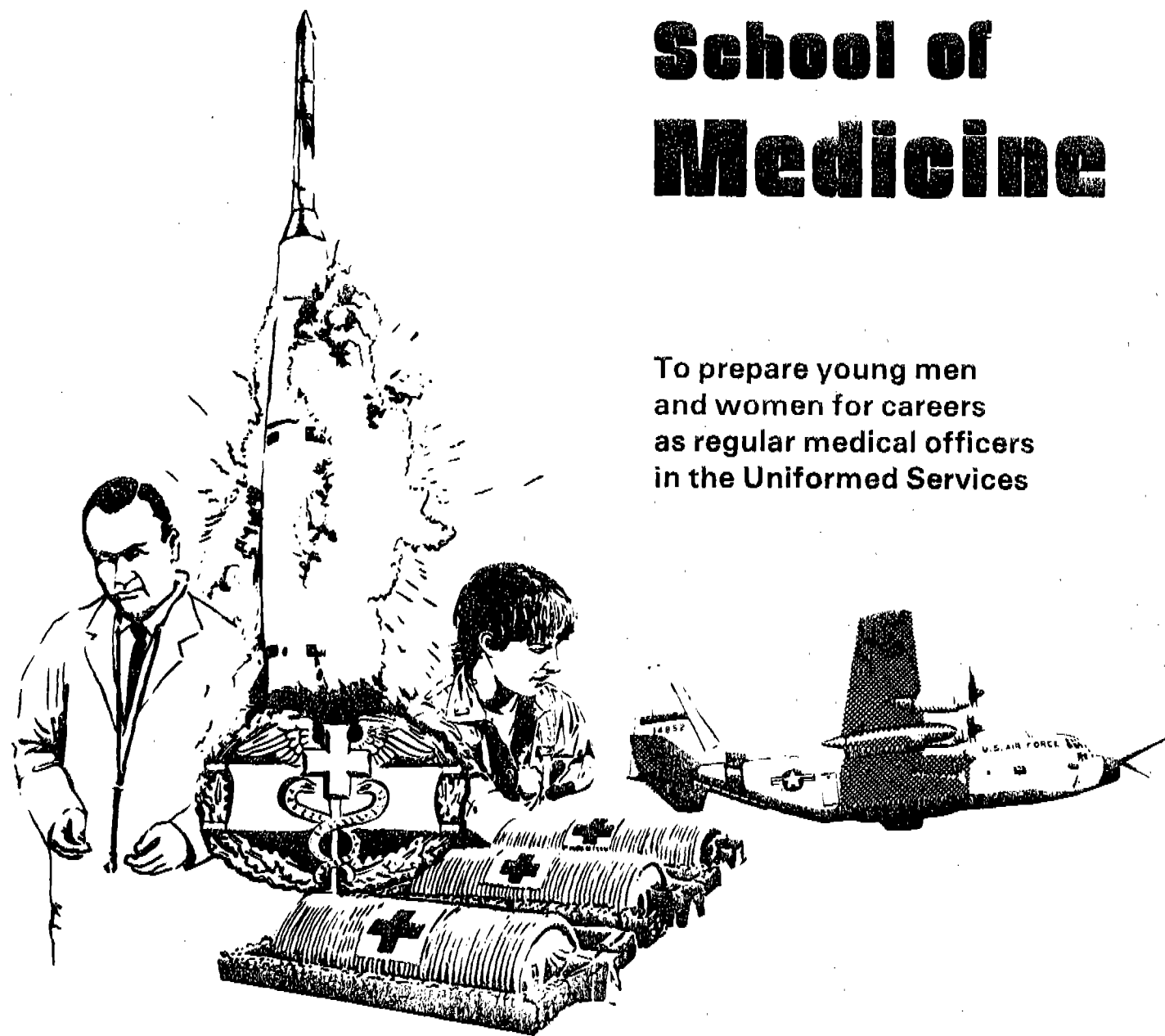
Admissions Office
Uniformed Services University
of the Health Sciences
4301 Jones Bridge Road
Bethesda, Maryland 20814
Tel: (301) 295-3101 or Autovon 295-3101

GPO : 1983 O - 399-386



The School of Medicine

To prepare young men
and women for careers
as regular medical officers
in the Uniformed Services



GENERAL INFORMATION

Created by public law in 1972, the USUHS was founded for the purpose of training young men and women for careers as health care professionals in the Uniformed Services.

In 1976, the University's School of Medicine admitted its first class of 32 freshman medical students. Sixty-eight medical students entered the Uniformed Services University of the Health Sciences (USUHS) in 1977; 108 in 1978; 124 in 1979, and the following year 130 students gained admittance. In 1981 and 1982, 156 medical students were admitted and by the mid-1980's, the School of Medicine projects a first-year class of 176 students, the planned enrollment capacity of the School.

The charter class studied for one year in interim facilities at the Armed Forces Institute of Pathology on the grounds of the Walter Reed Army Medical Center. In August 1977, the School of Medicine commenced its preclinical teaching activities at the University's new, permanent facilities on the grounds of the Naval Medical Command National Capital Region in suburban Bethesda, Maryland. Surrounded by master-planned communities, parks, and open land, the Center is adjacent to Interstate 495, a modern freeway system that circles the greater Washington area. The school's permanent campus occupies an area of more than one hundred acres.

Four connected buildings make up the permanent complex and were built at a total cost of approximately \$80 million. The facilities include staff and faculty offices, classrooms, student multidisciplinary laboratories, a lounge and cafeteria, student study areas, departmental laboratories, and academic support units such as a learning resources center, an electron microscopy suite, and a vivarium. Instructional and study areas are equipped with closed-circuit television.

The preponderance of clinical instruction for students is provided at the three major military medical centers in the Washington, D.C. area: Walter Reed Army Medical Center, Malcolm

Grow USAF Medical Center, and the Naval Hospital, Bethesda. Long recognized as being among the country's finest facilities for undergraduate and graduate medical education these centers have large outpatient populations have, collectively, more than 2,000 teaching beds and offer residencies in all of the major specialties. In addition, clinical experiences are scheduled for students at Wilford Hall USAF Medical Center in San Antonio, Texas; Naval Regional Medical Center, Jacksonville, Florida; Eisenhower Army Medical Center, Fort Gordon, Georgia; Naval Regional Medical Center, Charleston, South Carolina, and DeWitt Army Medical Center, Fort Belvoir, Virginia. The School operates in close association with other military medical facilities throughout the country and many other Federal health resources, such as the National Institutes of Health and National Library of Medicine, to provide a broad range of complementary preclinical and clinical experience for students.

CURRICULUM

The School of Medicine's four-year program which culminates in the award of the doctor of medicine degree, is aimed at: (1) developing students into competent, compassionate military physicians; (2) creating and fostering a learning environment that inspires investigative curiosity and the advancement of knowledge; and (3) providing a setting for the inculcation and furtherance of military medical professionalism.

The first two years of the curriculum consist predominately of preclinical instruction. The last two are devoted to the clinical disciplines. The integration between the clinical and basic sciences is progressive and proceeds with involvement in patient care activities early in the curriculum, starting with the first semester of the freshman year. While the overall program is designed to educate students to serve as providers of primary health care, there is sufficient flexibility in the curriculum to accommodate differences in interest among

students and also sufficient substance to enable graduates to pursue postgraduate activities such as research. Elective courses are offered in the clinical and research facilities of this country and also in areas of the world where diseases rarely seen in the United States are responsible for 80 percent of the morbidity and mortality. The curriculum also includes basic military orientation and concentration on unique aspects of military medicine.

GENERAL REQUIREMENTS FOR ADMISSION

Applicants must be citizens of the United States and must meet the physical and personal qualifications for a commission in the Uniformed Services. An applicant should not be older than 28 years of age as of 30 June of the year that he/she plans to enter the School of Medicine. A few waivers have been granted, but such exceptions are rare. A baccalaureate degree is required in addition to one year each of college English, general chemistry, organic chemistry, physics, general biology and mathematics. The New Medical College Admission Test (New MCAT) is also required of all applicants to the School of Medicine. Applicants must provide test scores that have been taken within three years of desired matriculation. The test is given in the fall and spring of each year. The spring testing may not be used for consideration if an individual wants to gain admittance to the first-year class beginning the same year, e.g., the spring, 1984 MCAT cannot be used by applicants who wish to enter the School of Medicine in July 1984, but the spring or fall 1984 test may be used by applicants who are applying for the 1985 first-year class. Information on registration for the MCAT is available from the American College Testing Program, Post Office Box 414, Iowa City, Iowa 52243 (telephone 319-337-1270).

Civilians and military personnel are eligible to apply. However, individuals who are in military

service or a program of study sponsored by the Armed Forces (including ROTC) must obtain a "Letter of Approval to Apply" from their respective service as part of their application. Each military department has established regulations governing the procedures for initiating and processing requests for approval. Inasmuch as the entering students will be commissioned officers in the military services they must, in addition to demonstrating the academic qualifications for the study of medicine, present evidence of a strong commitment to serving the United States as medical officers.

PROCEDURE FOR APPLICATION

The School of Medicine participates in the American Medical College Application Service (AMCAS). Application forms should be requested directly from AMCAS, 1776 Massachusetts Avenue, Northwest, Suite 301, Washington, D.C. 20036 (telephone 202-828-0600). *The School of Medicine does not distribute application packets.*

The School's Committee on Admissions will review all AMCAS applications and will decide on the basis of merit, taking into account both personal and intellectual characteristics, which individuals should advance to further stages of screening. Applicants should not send transcripts, letters of recommendation, or other materials until specifically requested to do so by the Admissions Committee. The Admissions Office will schedule personal interviews for those candidates that the Committee considers to be finalists in the screening process. The School of Medicine does not have any application fees; however, applicants are responsible for the AMCAS application fee and for incidental expenses such as postage and cost of travel for interview. Interviews are currently held at both the School of Medicine and regionally in San Francisco, California.

Applications must be submitted directly to AMCAS between 15 June and 1 November. Applicants are advised to submit all materials, including transcripts, to AMCAS well in advance of the deadline, as applications that are not complete and received by the 1 November deadline cannot be considered.

First-Year students are admitted only in July of each year. There are no provisions for transfer students and all students must enter at the first-year level.

GENERAL SELECTION FACTORS

Each year the University receives many more applications than the School of Medicine has positions to offer. Hence, placement in the class is on a competitive basis, decided by action of the Admissions Committee and the Dean, and granted only to the best qualified candidates in terms of demonstrated ability and potential for undertaking the study and practice of military medicine. The School of Medicine subscribes fully to the policy of equal educational opportunity. There are no quotas by race, sex, religion, marital status, national origin, socioeconomic background, or state of residence. There are no Congressional quotas or appointments.

Further, USUHS is committed to removing barriers that have made it more difficult for minorities, women, and economically disadvantaged college graduates to realize career goals in medicine and the military. To that end, the School of Medicine has established a program called "AQUA" in its admission office. AQUA stands for Accession of Qualified Under-represented Applicants. Through AQUA, the School seeks to identify and encourage applicants from groups which are under-represented in military medicine.

These categories include U.S. citizens who are women, black Americans, American Indians/Alaskan natives, Asian or Pacific Islanders,

Mexican Americans, Hispanics, and Puerto Ricans. AQUA also addresses college pre-med and science majors who have demonstrated motivation through ROTC program participation or prior active or reserve duty in the uniformed services.

For the 1982 freshman class 3,074 individuals applied. All new entrants had baccalaureate degrees, had taken the New MCAT and had been interviewed. The 156 matriculants had the following credentials: grade-point average, mean of 3.43; age at time of application, mean of 23.3; sex, 22 percent female; undergraduate major, 35 percent biology, with chemistry ranking second, and engineering (biomedical and mechanical), oceanography, nutrition, physics, business, psychology and physiology among the other disciplines represented; residence, 40.4 percent from northeastern states, 37.8 percent from western states, 15.4 percent from southern states, and 6.4 percent from central states.

MILITARY OBLIGATION

Upon entering the first-year class of the School of Medicine, students are commissioned and serve on active duty reserve in the grade of Second Lieutenant in either the Army or Air Force, or Ensign in the Navy or Public Health Service, receiving the appropriate pay and benefits of that grade. There are no tuition or fees for attending the School of Medicine. Required books, equipment, and instrument are also furnished without charge.

At graduation, upon the receipt of the doctor of medicine degree, students are promoted to the rank of Captain in the Army or Air Force, or Lieutenant in the Navy or Public Health Service. Graduates are obligated to serve on active duty as medical officers for not less than seven years. Periods of time spent in graduate medical education are not creditable toward satisfying this seven year obligation. A student who is dropped from the program, for either academic deficiencies or other reasons, may be required to perform active duty in an appropriate military capacity for a period equa

to the time spent in the program.

SERVICE BENEFITS

There are numerous advantages to a career in military medicine. The Uniformed Services, with their vast network of health resources including numerous hospital centers, research complexes, specialized educational and treatment facilities, and consultative agencies, provide physicians with opportunities for flexible career patterns, specialty training, and continued professional growth. Military medicine allows the practicing physician to work with highly trained, dedicated supporting staff and professionals, and to work with modern medical equipment and facilities in meeting the curative and preventive health care needs of the Services' members, their dependents, and the retired population. Moreover, military medicine is comprehensive, consisting not only of all of the customary specialties, but also a number of other sophisticated clinical fields such as aerospace, tropical, preventive, and nuclear medicine.

There are also a number of personal advantages associated with being a career Medical Corps officer. Currently, military physicians qualify for retirement after twenty years of active service. They do not have to contribute any part of their salary to retirement and do not have to invest or risk capital to ensure a retirement income in later years.

The salary schedule for military physicians is also competitive. By mid-career, most practicing military physicians earn in excess of \$40,000 annually in pay, allowances, and bonuses. While this may not compare with the gross income of physicians in private practice, the military physician does not have to pay overhead expenses such as rent, utilities, liability insurance, and payrolls. Hence, many civilian practitioners must earn considerably more to net as much as the mid-career military physician.

Opportunities for travel also make military life

exciting and attractive. An assignment abroad provides the military physicians and their families the occasion to become intimately acquainted with a foreign culture and people. The cost of moving expenses, whether stateside or overseas, is paid by the Services, and each of the military departments makes every effort to accommodate the assignment preferences of its physicians.

Comprehensive medical and dental care is provided by the Services for military physicians. Dependents of active-duty personnel are also entitled to medical treatment and care at facilities of the Uniformed Services on a space-available basis, or under certain circumstances, from a civilian medical resource at partial government reimbursement. Charges for other types of health care for dependents vary depending on circumstances, but are generally much lower than they would be under most other medical care programs.

Both abroad and in the United States, the Services offer a wide variety of recreational and social activities for military personnel and their families. Virtually all of the large, established military posts and bases have golf courses, gymnasiums, swimming pools, bowling lanes, tennis courts, theaters, craft shops, auto shops, riding clubs, gun clubs, teen clubs, and other recreational facilities. Officers' clubs offer a broad range of optional social activities for officers and their spouses.

Military physicians are eligible for thirty days of paid vacation annually. They also, while on active duty, are eligible for Serviceman's Group Life insurance, a term protection plan providing unrestricted coverage up to \$35,000 at a low annual premium.

Military doctors and their dependents are entitled to use commissary and post exchange facilities. In addition, they are entitled to professional advice and assistance without charge for a variety of problems of a personal nature (e.g., advice on income tax matters, the execution of personal wills, etc.).

In sum, the Uniformed Services offer physicians the time to concentrate on the challenges of medicine, and at the same time offer them a competitive salary, a secure financial future, and a welcome balance between professional duties and private life.

TABLE OF CONTENTS*

Title	<u>Page</u>
Foreword	iii
Table of Contents	xi
Program	xv
INJURY SEVERITY SCORING AND APPLICATIONS TO COMBAT CAUALTY CARE	
William J. Sacco and Howard R. Champion	1
A COMBINED BAYES-SAMPLING THEORY METHOD FOR MONITORING A BERNOULLI PROCESS	
Robert L. Launer and Nozer D. Singpurwalla	23
AN OPTIMAL SEQUENTIAL BERNOULLI SELECTION PROCEDURE	
Robert E. Bechhofer	27
A NEW METHOD OF EVALUATING NORMAL AND t -TAIL AREAS	
Andrew P. Soms	49
THE DESIGN OF A QUANTAL RESPONCE EXPERIMENT: AN EMPIRICAL APPROACH	
Refik Soyer	55
INFORMATIVE QUANTILE FUNCTIONS AND IDENTIFICATION OF PROBABILITY DICTRIBUTION TYPES	
Emanual Parzen	97
ON THE LEHMANN POWER ANALYSIS FOR THE WILCOXON RANK SUM TEST	
James R. Knaub, Jr.	107

*This table of contents contains only the papers that are published in this technical manual. For a list of all papers presented at the twenty-Ninth Conference on the Design of Experiments, see the Program of this meeting

COMPLEX DEMODULATION - A TECHNIQUE FOR ASSESSING PERIODIC
COMPONENTS IN SEQUENTIALLY SAMPLED DATA

Helen C. Sing, Sander G. Genser, Harvey Babkoff, David R. Thorne,
and Frederick W. Hegge 131

CYCLES OF SUICIDE

Joseph M. Rothberg 157

EVALUATION OF OPTICAL DATA COLLECTION INSTRUMENTATION IN THE DESERT
ENVIRONMENT

Robert A. Dragon 167

A TYPE OF CORRELATED DATA IN OPERATIONAL TESTING

Ellen Hertz 173

A SIMULATION PROCESS FOR DETERMINING RELIABILITY OF CYCLIC RANDOM
LOADED STRUCTURES

D. Neal, W. Matthews, and T. DeAngelis 177

RANDOM NUMBERS FROM SMALL CALCULATORS

Donald W. Rankin 203

APPLICATION OF THE BOOTSTRAP METHOD TO A MEASURE OF FORCE
EFFECTIVENESS (AN EMPIRICAL CASE STUDY)

Eugene Dutoit, Ellen Shannahan, and Joseph Tessmer 219

ACCEPTANCE OF A MEAL AND ITS COMPONENTS - AN EXERCISE IN MISSING DATA

Edward W. Ross 235

NUMERICAL VALIDATION OF TUKEY'S CRITERION FOR CLINICAL TRIALS AND
SEQUENTIAL TESTING

Charles R. Leake 259

FIRE SUPPORT TEAM EXPERIMENT

Jock O. Grynoviecki, Jill H. Smith, Virginia A. Kaste, and
Ann E. McKaig 263

A TECHNIQUE TO APPROXIMATE COMPLEX COMPUTER MODELS - AN APPROXIMATION
OF THE TEISBERG MODEL

Joseph Tessmer 307

HIGH TO LOW DOSE EXTRAPOLATION OF EXPERIMENTAL ANIMAL
CARCINOGENESIS STUDIES

Charles C. Brown	329
ATTENDANCE LIST	355

A G E N D A

for the

TWENTY-NINTH CONFERENCE ON THE DESIGN OF EXPERIMENTS IN

ARMY RESEARCH, DEVELOPMENT AND TESTING

19-21 October 1983

Host: Uniformed Services University of Health Sciences

Location: Bethesda, Maryland

***** Wednesday, 19 October *****

0815-0915 REGISTRATION

0915-0930 CALLING OF THE CONFERENCE TO ORDER

David F. Cruess, Department of Preventive Medicine & Biometrics,
Uniformed Services University of Health Sciences

WELCOMING REMARKS

0930-1200 GENERAL SESSION I (Auditorium, Bldg. B)

Chairman: David F. Cruess, Uniformed Services University of
Health Sciences

0930-1030 KEYNOTE ADDRESS

EPIDEMIOLOGY AND RISK ASSESSMENT: COURTS, CLOCKS AND CONFUSION

Marvin A. Schneiderman, National Cancer Institute, Bethesda,
Maryland

1030-1100 BREAK

1100-1200 INJURY SEVERITY SCORES AND APPLICATIONS TO MILITARY TRIAGE

William Sacco, Bellaire, Maryland

1200-1315 LUNCH

1300-1715 SPECIAL SESSION ON SEQUENTIAL TESTING (Bldg. A, Room C)
Chairman: Michael Woodroffe, University of Michigan and Rutgers University

1300-1330 RELIABILITY MONITORING WITH BERNOULLI SAMPLING
Daniel Willard, Office of the Deputy Under Secretary of the Army

1330-1415 A COMBINED BAYES SAMPLING THEORY APPROACH TO TRUNCATED SEQUENTIAL BERNOULLI TESTING
Robert L. Launer, U. S. Army Research Office and Nozer D. Singpurwalla, George Washington University

1415-1445 SENSITIVITY TESTING IN BALLISTICS
J. Richard Moore, Ballistic Research Laboratory

1445-1515 BREAK

1515-1545 A TRUNCATED SEQUENTIAL PROBABILITY RATIO TEST
J. Richard Moore, Ballistic Research Laboratory

1545-1630 EFFICIENT SEQUENTIAL DESIGNS FOR SENSITIVITY EXPERIMENTS
C. F. Wu, University of Wisconsin-Madison

1630-1715 A SEQUENTIAL BERNOULLI SELECTION PROCEDURE
Robert E. Bechhofer, Cornell University

1830-1930 ***** CASH BAR AT OFFICER'S CLUB, *****
NAVAL MEDICAL CENTER

1930- ***** BANQUET AND PRESENTATION OF WILKS AWARD, *****
OFFICER'S CLUB, NAVAL MEDICAL CENTER

***** Thursday, 20 October *****

0830-1030 **TECHNICAL SESSION I - "Statistical Theory"**
 (Bldg. A, Room C)

Chairman: Malcolm Taylor, Ballistic Research Laboratory

MODEL IDENTIFICATION OF PROBABILITY DISTRIBUTIONS USING INFORMATIVE QUANTILE FUNCTIONS

Emanuel Parzen, Texas A&M University

ON THE LEHMANN POWER ANALYSIS FOR THE WILCOXON RANK SUM TEST

James R. Knaub, Jr., US Army Logistics Center

A NEW METHOD OF CALCULATING NORMAL AND t TAIL PROBABILITIES

Andrew P. Soms, University of Wisconsin-Madison

0830-1030 TECHNICAL SESSION II - "Analysis of Longitudinal Data"
(Bldg. A, Room B)

Chairman: Charles R. Leake, US Army Concepts Analysis Agency

COMPLEX DEMODULATION - A TECHNIQUE FOR ASSESSING PERIODIC COMPONENTS IN SEQUENTIALLY SAMPLED DATA

Helen C. Sing, Sander G. Genser, Harvey Babkoff, David R. Thorne
and Frederick W. Hegge, Walter Reed Army Institute of Research.

HOW GOOD ARE TRAJECTORY ERROR ESTIMATES?

William S. Agee and Robert H. Turner, White Sands Missile Range

1030-1100 BREAK

1100-1200 GENERAL SESSION II (Auditorium, Bldg. B)

Chairman: James J. Schlesselman, Uniformed Services University
of Health Sciences

TITLE TO BE ANNOUNCED: INTERACTIVE DATA ANALYSIS

Jerome Friedman, Stanford Linear Accelerator Center

1200-1330 LUNCH

1330-1600

CLINICAL SESSION (Bldg. A, Room C)

Chairman: Carl Bates, US Army Concepts Analysis Agency

Panelists: Robert E. Bechhofer, Cornell University

Charles Brown, National Cancer Institute

Churchill Eisenhart, National Bureau of Standards

Dennis E. Smith, Desmatics, Inc.

Andrew P. Soms, University of Wisconsin - Madison

Chien Fu Wu, University of Wisconsin - Madison

CYCLES OF SUICIDE

Joseph M. Rothberg, Walter Reed Army Institute of Research

EXPERIMENT TO DETERMINE EVALUATION OF CURRENT DATA COLLECTION
OPTICAL INSTRUMENTATION IN THE DESERT ENVIRONMENT

Robert A. Dragon, White Sands Missile Range

A TYPE OF CORRELATED DATA IN OPERATIONAL TESTING

Ellen Hertz, US Army Operational Test and Evaluation Agency

1330-1500

TECHNICAL SESSION III - "Simulation Techniques and Applications"
(Bldg. A, Room B)

Chairman: William Baker, Ballistic Research Laboratory

A SIMULATION PROCESS FOR DETERMINING RELIABILITY OF FATIGUE
LOADED STRUCTURES

Donald M. Neal, US Army Materials and Mechanics Research Center

RANDOM NUMBERS FROM SMALL CALCULATORS

Donald W. Rankin, White Sands Missile Range

APPLICATION OF THE BOOTSTRAP METHOD TO A MEASURE OF FORCE
EFFECTIVENESS

Eugene F. Dutoit and Ellen Shannahan, US Army Infantry School

1500-1530

BREAK

1530-1700

TECHNICAL SESSION IV - "Test and Evaluation Techniques"
(Bldg. A, Room B)

Chairman: Langhorne Withers, US Army Operational Test and
Evaluation Agency

LIFE OF AIRCRAFT BEARINGS RESTORED THROUGH GRINDING OF RACEWAYS

Martin W. Joseph, US Army Troop Support and Aviation Materiel
Readiness Command, and Harold Schuetz, Stew Chen and Mitten
Dutta, US Army Aviation R&D Command, St. Louis

ACCEPTANCE OF A MEAL AND ITS COMPONENTS - AN EXERCISE IN MISSING
DATA

Edward W. Ross, Jr., US Army Natick Research and Development
Laboratories

NUMERICAL VALIDATION OF TUKEY'S CRITERIA FOR CLINICAL TRIALS AND
SEQUENTIAL TESTING

Charles R. Leake, US Army Concepts Analysis Agency

***** Friday, 21 October *****

0830-1000

TECHNICAL SESSION V - "Application in Experimental Design"
(Bldg. A, Room C)

Chairman: Jerry Thomas, Ballistic Research Laboratory

APPLICATION OF FACTORIAL ANOVA PROCEDURES TO FABRIC TESTING
PROBLEM

Raymond V. Spring, US Army Natick Research and Development
Laboratories

A TECHNIQUE TO APPROXIMATE COMPLEX COMPUTER MODELS

Joseph M. Tessmer, Department of Energy, Office of the Strategic
Petroleum Reserve

FIRE SUPPORT TEAM HEADQUARTER EXPERIMENT

Jock O. Grynovicki and Jill H. Smith, Ballistic Research
Laboratory

1000-1030

BREAK

1030-1200

GENERAL SESSION III - (Auditorium, Bldg. B)

Chairman: Douglas B. Tang, Walter Reed Army Institute of
Research; AMSC Subcommittee on Probability and
Statistics

OPEN MEETING of the Subcommittee on Probability and Statistics

HIGH TO LOW DOSE EXTRAPOLATION OF EXPERIMENTAL ANIMAL
CARCINOGENESIS STUDIES

Charles Brown, National Cancer Institute

1200

ADJOURN

INJURY SEVERITY SCORING AND APPLICATIONS TO COMBAT CASUALTY CARE

William J. Sacco

Howard R. Champion

Washington Hospital Center

PREFACE

Injury Scales have wide applications to management of trauma victims in civilian and military settings. They are used for epidemiological studies, prediction of outcome, triage and monitoring, assessment of clinical modalities, and for evaluation of patient management.

This paper is a product of over ten years of research by the authors toward developing and validating indices that measure injury severity.¹⁻¹³

The research began in 1972 at the Maryland Institute for Emergency Medical Systems (MIEMS) and the Aberdeen Proving Ground with support from the Department of Army, and since 1976 has continued at the Washington Hospital Center (WHC), Washington, D. C., with support largely from the Department of Health and Human Services and the Department of the Navy. A number of severity indices were developed and tested on a computerized data base of over 5,000 patients seen at WHC. Methods for developing and validating indices were refined, and methods of triage, monitoring, and evaluation of care were developed that used severity indices to describe the patient population in terms of degree of injury and probability of survival. The indices are based on easily attained data and have proved to be reliable predictors of outcome in a number of trauma centers.

In this paper we describe three indices and military applications.

The indices are the Injury Severity Score,¹⁴ the Trauma Score,⁹ and the Global Index.^{15,16} The Injury Severity Score is based on injury descriptions in terms of anatomical lesions. The Trauma Score is based on assessments of physiological responses soon after injury. The Global Index characterizes patient condition in the intensive care unit using measures of organ function.

Injury Severity Score

The Injury Severity Score is based on the Abbreviated Injury Scale¹⁷ (AIS), another well known anatomical scale. The AIS relies on a list of lesions. Each lesion is assigned a severity code from 1 (for minor injuries) to 6 (for injuries that are untreatable and always fatal). Thus, the characterization of a multiply injured patient in terms of AIS would consist of a string of numerical codes.

The ISS is based on AIS severity codes for six body regions, head and neck, face, chest, abdominal and pelvic contents, extremities and pelvic girdle, and external.

The ISS ranges from 1 to 75. The higher the score, the poorer the patient's condition. If a victim has any injury with an AIS value of 6 the ISS is assigned a value of 75. Otherwise,

to compute ISS one first identifies the highest AIS code in each of the six body regions, and the squares of the highest three of the six codes are added to obtain the ISS.

Trauma Score

The Trauma Score is a physiological measure of injury severity. It is based on seven circulatory, respiratory and neurological assessments easily obtained by doctors, nurses, or paramedics.

The Trauma Score is developed from these assessments as shown in Table 1. Eye opening, best verbal response, and best motor response make up the Glasgow Coma Scale,¹⁸ which is used worldwide to assess central nervous system function.

TABLE 1

TRAUMA SCORE
CATEGORY DEFINITIONS, METHODS OF ASSESSMENT, AND CODES

		<u>Rate</u>	<u>Codes</u>	<u>Score</u>
A.	<u>Respiratory Rate</u>	10-24	4	
	Number of respirations in 15 seconds;	25-35	3	
	multiply by four	36 or greater	2	
		1- 9	1	
		0	0	A. ____
B.	<u>Respiratory Expansion</u>			
	<u>Normal</u>	Normal	1	
	<u>Retractive</u> - Use of accessory muscles	Retractive	0	B. ____
C.	<u>Systolic Blood Pressure</u>	90 or greater	4	
	<u>Systolic cuff pressure</u> - either arm,	70-89	3	
	by auscultation or palpation	50-69	2	
		1-49	1	
	No pulse	0	0	C. ____
D.	<u>Capillary Refill</u>			
	<u>Normal</u> - Nail bed color refill			
	in 2 seconds	Normal	2	
	<u>Delayed</u> - More than 2 seconds capillary refill	Delayed	1	
	<u>None</u> - No capillary refill	None	0	D. ____
E.	<u>Glasgow Coma Scale</u>			
		<u>Total</u>		
		<u>GCS Points</u>	<u>Score</u>	
1.	<u>Eye Opening</u>			
	Spontaneous	4	14-15	5
	To Voice	3	11-13	4
	To Pain	2	8-10	3
	None	1	5- 7	2
			3- 4	1
				E. ____
2.	<u>Best Verbal Response</u>			
	Oriented	5		
	Confused	4		
	Inappropriate Words	3		
	Incomprehensible Sounds	2		
	None	1		
3.	<u>Best Motor Response</u>			
	Obeys Commands	6		
	Localizes Pain	5		
	Withdraw (pain)	4		
	Flexion (pain)	3		
	Extension (pain)	2		
	None	1		

Total GCS Point (1+2+3) _____

TRAUMA SCORE _____
(Total Points A+B+C+D+E)

To illustrate computation of the Trauma Score, an example is given below for a hypothetical patient:

Assessment	Result	Score
Respiratory rate	3 in 15 seconds	<u>4</u>
Respiratory expansion	Normal	<u>1</u>
Systolic blood pressure	127	<u>4</u>
Capillary refill	Normal	<u>2</u>
Glasgow Coma Scale		
Eye opening	Spontaneous (4)	
Best verbal response	Oriented (5)	
Best motor response	Obeys commands (6)	
	Total GCS = 15	<u>5</u>
Trauma Score =		<u>16</u>

Table 2 contains probabilities of survival for values of the Trauma Score based on penetrating injury data.¹⁹

TABLE 2. Probabilities of Survival, P_S , for each value of the Trauma Score

<u>TS</u>	<u>Probability of Survival</u>
16	0.99
15	0.98
14	0.97
13	0.94
12	0.89
11	0.82
10	0.70
9	0.55
8	0.40
7	0.26
6	0.15
5	0.088
4	0.048
3	0.026
2	0.014
1	0.007

Global Index

The Global Index is used in the intensive care unit to characterize patient condition:

$$\text{Global Index} = R_n + C_n + B_n + G_n,$$

where

$$R_n = 1.5 \times \text{Respiratory Index}$$

$$C_n = 0 \text{ if serum creatinine is one or less}$$

$$2.0 \times (\text{serum creatinine} - 1.0) \text{ otherwise}$$

$$B_n = 0.5 \times \text{serum bilirubin}$$

$$G_n = 15.0 - \text{Glasgow Coma Scale}$$

The Respiratory Index (RI), a measure of respiratory insufficiency, is defined as follows:

$$RI = \frac{713 F_{I O_2} - P_a CO_2 - P_a O_2}{P_a O_2}$$

where:

$F_{I O_2}$ = fractional concentration of O_2 in
inspired gas

$P_a O_2$ = arterial partial pressure of oxygen in torr

$P_a CO_2$ = arterial partial pressure of carbon dioxide in torr

The numerator of the RI is an approximation of the alveolar-arterial oxygen difference, which is an indicator of oxygen sufficiency and an important consideration in controlling arterial oxygenation.

The RI, serum creatinine, serum bilirubin, and the Glasgow Coma Scale have proven to be excellent indicators of renal, hepatic, and central nervous system function, respectively, in trauma patients.^{1-6,10,11}

Application to management of Combat Casualties

Here we discuss applications of the indices to the triage, tracking, and evaluation of management of casualties.

Triage Principles Incorporating a Physiological Response Score

Triage is a method of managing mass casualties including assessment and classification of casualties for priorities of treatment and evacuation. In a wartime mass casualty situation, the priorities of treatment and evacuation are dependent obviously on military objectives. The priorities can be radically different for different objectives.

The triage principles discussed here, which implement physiological response scores, are intended to maximize survivors. As such, these principles would be appropriate after other higher priority objectives (if any) had been addressed.

By definition, in a mass casualty situation, resources are not available for meeting the needs of all casualties over a short period of time. Hence triage is used to sequence patient care. If the objective is to maximize survivors, establishing urgency is the first sorting criterion.

The battalion aid station is the primary site of casualty sorting. Under some current military protocols, casualties are examined by the battalion aid station medical officer or assistants. The medical officer determines the level of treatment required and the priority of evacuation.

All casualties are classified by level of treatment required. There are four classification groups, called minimal, delayed, immediate, or expectant; defined as follows:

- 1) Minimal: Those casualties whose injuries are so slight that they can be managed by self-help or buddy care and can be returned promptly to their units for full duty.

- 2) Delayed: Those casualties whose wounds require medical care but are so slight that they can be managed by the battalion aid station or in the amphibian objective area and can be returned to duty after being held for only a brief period.
- 3) Immediate: Those casualties whose conditions indicate the need for immediate resuscitation, and usually surgery.
- 4) Expectant: Those casualties that have low chances of survival even if accorded full medical resources.

Triage in the field involves priorities for care in the field and for evacuation to higher echelons of care. Casualties may be triaged many times in the field. Frequency will depend upon such factors as the intensity of combat and availability of time and resources for resuscitation, treatment, or evacuation, or for more definitive assessment and treatment.

In such circumstances, serial measurements of a physiological response score can help provide a finer discrimination of patients in Categories 3 and 4 at various stages of triage and care. For example:

1. Each patient in Categories 3 and 4 can be assigned a probability of survival, P_g , associated with the response score. The P_g is to be interpreted as the probability of survival presuming immediate definitive care.

2. Serial assessments can be used to measure the clinical "change of state" of a casualty:

- a. From scene of wounding to Battalion Aid Station (BAS).
- b. Awaiting resuscitation therapy at the BAS.
- c. Before and after resuscitation at the BAS.
- d. In the holding area at or near the BAS.
- e. During evacuation.
- f. Awaiting additional care in the field hospital.

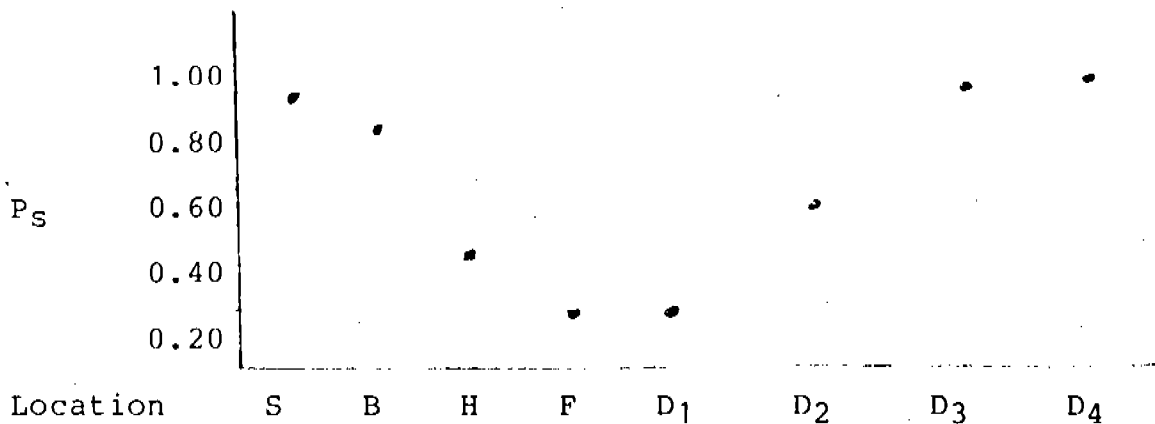
The serial scores would provide evidence of casualty deterioration, stability, or improvement.

Several studies have evaluated the Trauma Score as an adjunct to casualty triage in the early stages of combat care.^{20,21} The results showed that Navy Corpsmen were capable of obtaining Trauma Score assessments with minimal training and their facility and accuracy improved with repetitive drills and practice on simulated casualties.

Patient Tracking

The Trauma Score and Global Index can provide a permanent record of patient condition transitions from the injury scene through the ICU, with implications for triage, evaluation of care in general, and evaluation of specific therapeutic modalities at all echelons of care.

One of the simplest methods for tracking the progress of a casualty is a time series plot of the survival probability P_S , illustrated for a hypothetical patient in the figure below,



The symbols on the horizontal axis are defined as follows:

S: Scene of injury

B: Battalion Aid Station

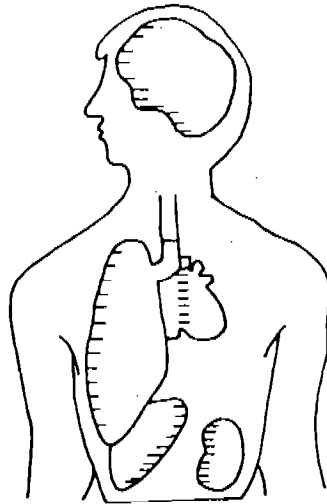
H: Holding Area

F: Field Hospital

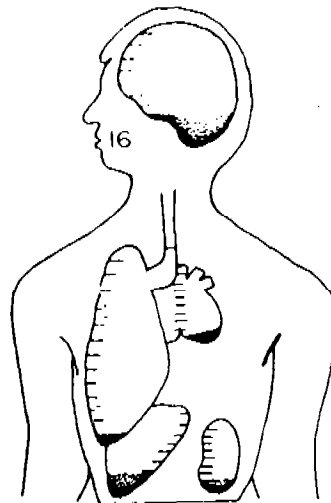
D_i: Definitive Care Facility (Admission and three succeeding days)

In the construction of such a chart, the probability of survival estimates for S, B, H, and F are based on a simple score (Trauma Score or variant) and those for D_i are based on the Global Index.

In addition, we can provide a graphical presentation of the ICU record by means of "anatoglyphs", like the diagram shown below.^{22,23}



In these anatoglyphs, the five body regions of greatest physiological importance (the brain, heart*, kidney, lungs, and liver) are outlined with scale markings. Shading these five areas to a height corresponding to the severity of the individual organ's derangement gives an anatoglyph of the patient's condition. An example is shown below.



*Although the heart is included here, this version of the Global Index does not contain cardiovascular variables.

The number appearing near the mouth of the profile is the Global Index. A series of daily anatoglyphs transforms the patient's charts into a picture sequence that can be read at a glance.

Evaluation of Care

Here we present a two-phase approach^{15,24} to evaluation of patient care.

The first level, called PRE (from PREliminary), identifies unexpected survivors and deaths. These cases may be therapeutic triumphs or failures. PRE can be used to assess patient management at any echelon.

The second level, the State Transition Screen or STS, identifies patients with unusual clinical courses in the definitive care unit. Among these are patients who improve substantially before they die, and patients who deteriorate substantially before they recover. These cases may be near triumphs or near failures.

PRE: Semi-Quantitative Assessment of Trauma Care

Ideally the basic ingredients of the PRE methodology are two injury severity scales, one anatomical, the other physiological. The goal of PRE is to identify cases where the outcome was anomalous -- in terms of the scales employed.

In the discussion here, we use the Trauma Score as the physiological assessment and the Injury Severity Score as the anatomical assessment.

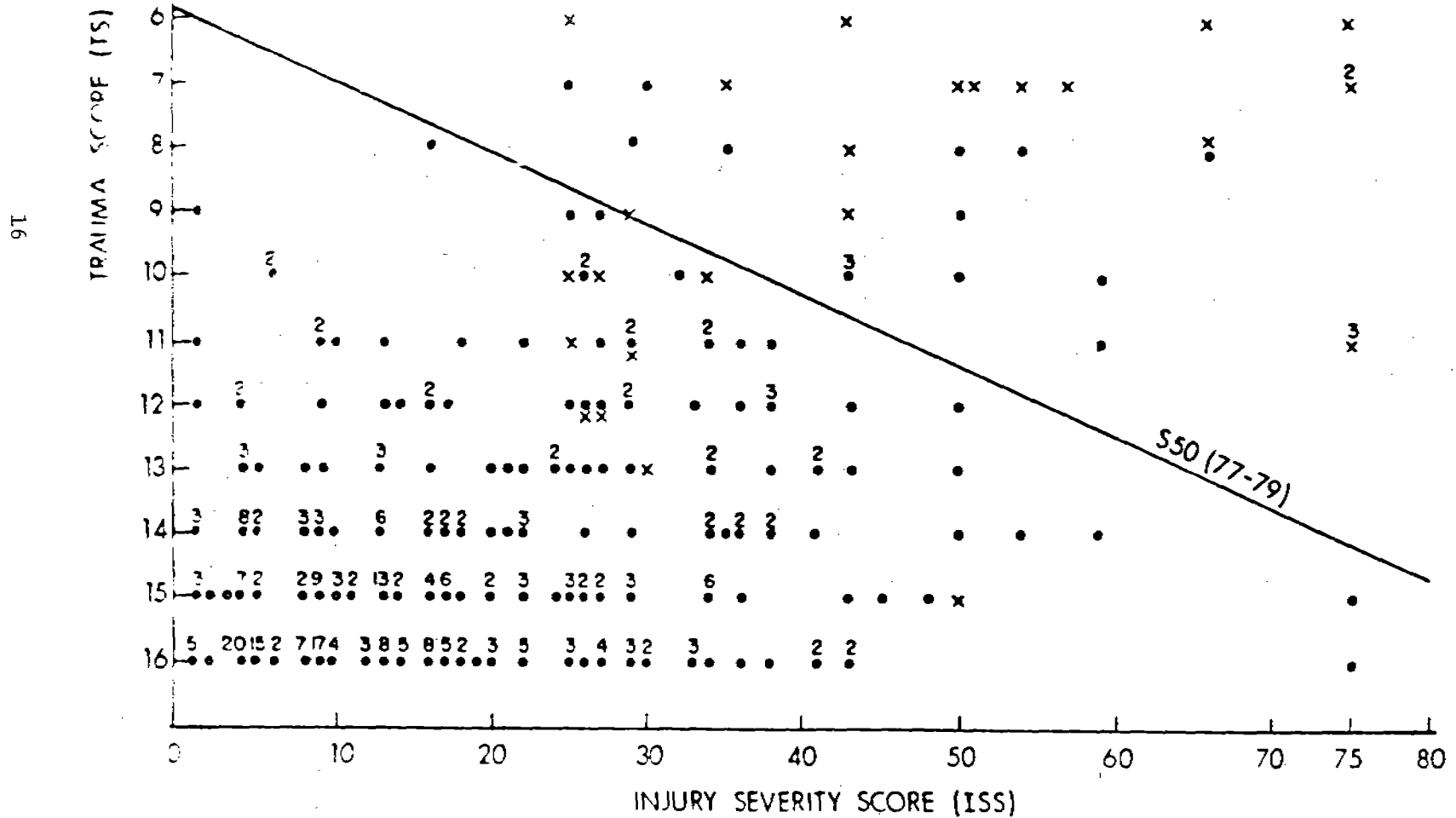
The scores are plotted on an x-y graph as in Figure 1. For example, a patient with an ISS of 25 and a TS of 13 is represented by an x or a dot at coordinates 25,13. The dots are survivors and the x's, deaths. Multiple occurrences at the same coordinates are indicated by a number near the symbol.

On such a plot, whatever the scales employed, survivors usually predominate toward one corner of the plot, deaths at the opposite corner; and mixed results are seen along a sloping line that cuts across the connecting diagonal. Such is the case in Figure 1, where survivors predominate at the lower left and deaths at the upper right. The sloping line in Figure 1 is called the S50 isobar. At each point on this line, the patient has a 50 percent chance of survival. A patient whose point is below the line in this figure has better than a 50 percent chance of survival, and in a statistical sense, is expected to survive.

The survivors whose points are above the line, and the nonsurvivors below the line, are the patients sought to be identified by PRE: those with anomalous or "unexpected" outcomes. These are cases worthy of audit.

The data in Figure 1 are from a set of 402 blunt trauma patients seen at the Washington Hospital Center (Washington,

Washington Hospital Center Shock Trauma Patients 1980-1981 BLUNT TRAUMA PATIENTS



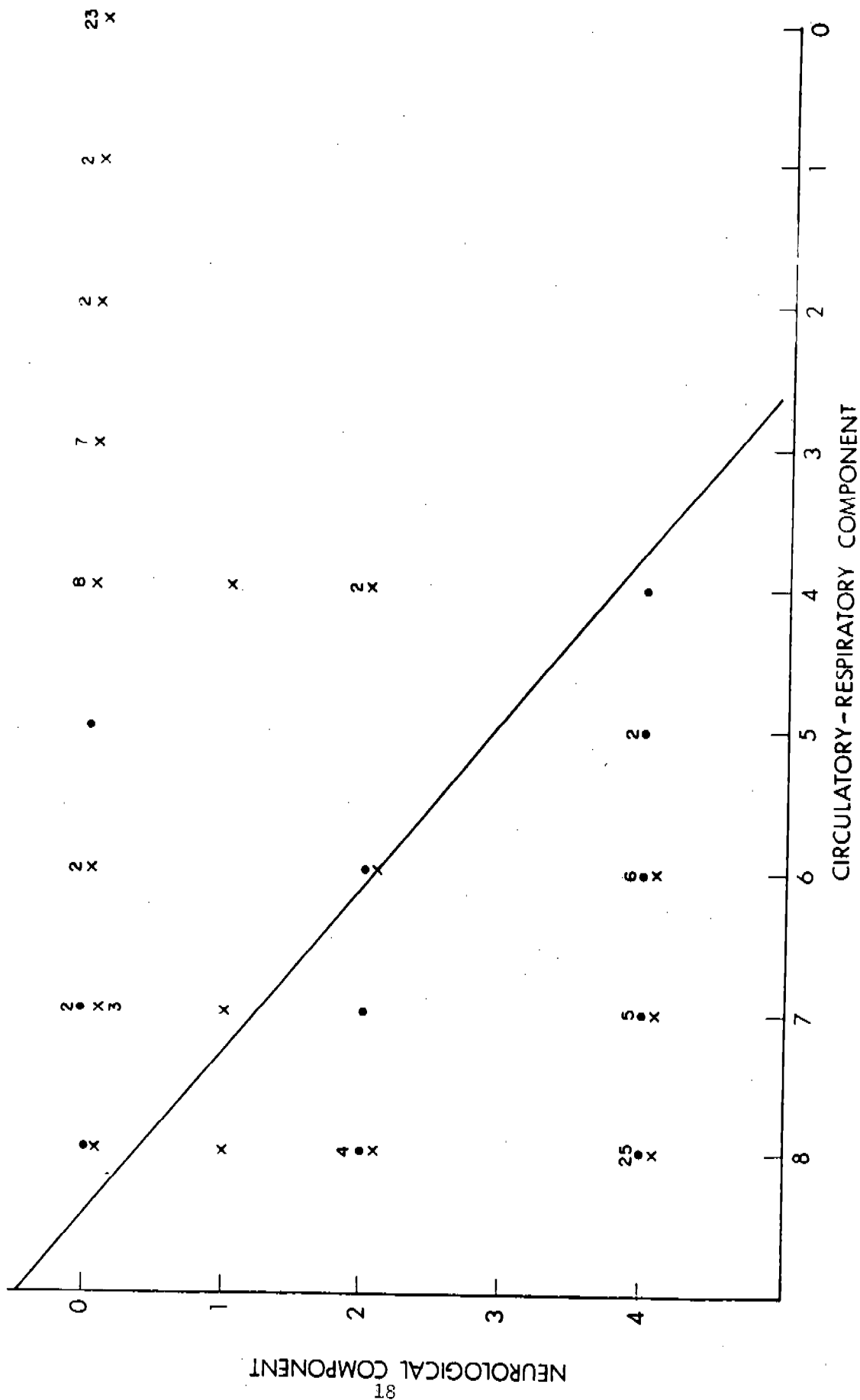
Trauma Scores versus Injury Severity Scores
Figure 1

D.C.) from January 1, 1980 to December 31, 1981.²⁴ The S50 isobar in the figure was computed from earlier data. This combination -- of current data and historic isobar -- illustrates the usual implementation of PRE. In practice, a patient's (ISS, TS) pair is plotted as soon as his data are available, and the decision whether the patient outcome was unexpected (in a statistical sense) is based on an isobar from previous data. PRE can also be implemented with two-component physiological scores. These pairs are nearly as powerful as the physiological anatomical pairs. The patient can be represented as soon as the measures are obtained. One need not wait for an anatomical assessment. Figure 2 is an example for a two-component pair applied to serious head injured patients.

State Transition Screen (STS)

The cases cited by PRE are not the only ones that are interesting and deserving of audit. Other interesting cases are those whose admission scores to definitive care facility indicate a better than 50 percent chance of survival, but who deteriorate substantially before they recover; and those whose admission scores indicate a low probability of survival, but who improve substantially before they die. To sift out these cases, we need measures of patient condition, and criteria for distinguishing major from minor fluctuations.

Serious Head Injury Patients



Physiological Pair of Assessments

Figure 2

The survival probabilities needed are the admission value (P_A) and daily values in the ICU. The admission value can be based on a Trauma Score - ISS combination or a two-component physiological score, and the ICU values can be based on the Global Index.

The audit selection criteria in STS are different for survivors and non-survivors. The survivors selected are those for whom P_A is greater than 0.50, but whose survival probability falls below P_A by 0.25 or more during the ICU stay. The non-survivors selected are those for whom P_A is 0.50 or less, but whose Global Index reaches 10 or less during the ICU stay.

References

1. Sacco W, Champion H, Nyikos P, et al. A renal index for multiple trauma. Edgewood Arsenal Technical Report EB-TR-74038. Aberdeen Proving Grounds, Aberdeen, MD, 1974.
2. Champion H, Sacco W, Long W, et al. Indications for early hemodialysis in multiple trauma. Lancet 8: 1125-1127, 1974.
3. Goldfarb M, Ciurej T, Sacco W, Weinstein M. A simple respiratory guide for respiratory therapy in the trauma patient. Edgewood Arsenal Technical Report EB-TR-73061. Aberdeen Proving Grounds, Aberdeen, MD 1974.
4. Goldfarb M, Sacco W, Cuirej, T. Tracking respiratory therapy in the trauma patient. Am J Surg 129:255-258, 1975.
5. Goldfarb M, Sacco W, Weinstein M, Ciurej T. Two prognostic indices for the trauma patient. Comput Biol Med 7:21-25, 1977.
6. Sacco W, Milholland A, Cowley R, et al. Trauma indices. Comput Biol Med 7:9-20, 1977.
7. Champion H, Sacco W, Hannan D, et al. Assessment of injury severity: The Triage Index. Crit Care Med. 8:201-208, 1980.
8. Sacco W, Champion H, Carnazzo A. Trauma Score. Current Concepts in Trauma Care. Spring 1981:9-11.
9. Champion H, Sacco W, Carnazzo A, Copes W, Fouty W. Trauma Score. Crit Care Med 9:672-676, 1981.
10. Champion H. Final report: Quantitation of injury and critical illness. National Center for Health Services Research Grant No. HS-02559, 1981.
11. Sacco W, Champion H, McLaughlin S. Use of glyphs in tracking patients. Proceedings of the Fourteenth Hawaii International Conference on System Sciences, Honolulu, 1981.
12. Champion H, Sacco W, Ashman W. An anatomical index of injury severity. Trauma 20:197-202, 1980.
13. Champion H, Sacco W. Trauma risk assessment: Review of severity scales. Emergency Medical Manual, Appleton-Century-Crofts, 1983.
14. Baker S, O'Neill B, Haddon W, Long W. Injury Severity Score: A method for describing patients with multiple injuries and evaluating emergency care. Trauma 14:187-196, 1974.

15. Champion H, Sacco W, Hunt T. Trauma severity scoring to predict mortality. World J Surg. 7:4-11, 1983.
16. Champion H, Sacco W, Lawnick, M. A simple global index for the evaluation of the care of trauma patients in the intensive care unit. In preparation.
17. Committee on Medical Aspects of Automotive Safety. Rating the severity of tissue damage: I. The Abbreviated Injury Scale. JAMA, 215:277-280, 1971.
18. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet 2:81, 1974.
19. Sacco W, Champion H, Gainer P, et al. Trauma Score for penetrating injuries. Submitted for publication.
20. Sacco W, Champion H, et. al. Testing of navy corpsmen in trauma score assessments. Proceedings of the Sixteenth Hawaii International Conference on System Sciences, 1983.
21. Sacco W, Champion H. Field testing of the trauma score. Final report for the Naval Medical Research and Development Command (#N00014-82-C-0577) March, 1983.
22. Sacco W, Champion H, et. al. Quantitative glyph characterization of patient state in the intensive care unit. Critical Care Medicine Symposium, St. Louis, MO, June, 1982.
23. Sacco W, et. al. Glyphs - getting the picture. The Science Teacher, February, 1983.
24. Sacco W, Champion H. A simple method for evaluating care of trauma patients. Current Concepts in Trauma Care, Spring, 1983.

A Combined Bayes-Sampling Theory Method For Monitoring a Bernoulli Process

Robert L. Launer, U.S. Army Research Office
Nozer D. Singpurwalla, George Washington University

We assume a population of one-shot missiles which are stored in a ready or near ready state at the physical point of their deployment. We hope that the missiles will sit in idle waiting for many years, but this allows environmental effects to degrade the missiles' capability of successful deployment. Since even a brand new missile may fail to operate properly, and there are no important physical differences between the individual missiles in the given population, we shall assume that a randomly selected missile will have a probability p_t of successful deployment, or reliability, at time t .

It is obviously important to monitor p_t , so a sample of the missiles is tested periodically. Since the testing is destructive, the population is eventually depleted by the testing. Furthermore, defects in missile design may be uncovered, so modifications may be introduced which will have a tendency to increase the reliability. For technical reasons, however, we choose to describe a test which is designed to detect a deterioration in the reliability.

No target value for the reliability is given by management, so that the testing at time t is used to determine if there has been a change in the reliability since time $t-1$. The following requirements are given and will be used to formalize a test of hypothesis to accomplish the goals of the testing procedure.

It is required to:

1. detect whether p_t has changed by an amount d^* since the immediately preceding testing period, with a probability of at least π at time $t=2,3,4,\dots$
2. compensate for the sampling uncertainty in $\hat{p}_t$, the estimate of p_t , in constructing the test of hypothesis.
3. use the minimum possible sample sizes in accomplishing requirements 1 and 2, above.

Since the test data are pass-fail in nature, the binomial probability model is appropriate for describing the stochastic sample behaviour. Suppose we choose the test size to be α for the hypotheses:

$$H_0: p_t = p_{t-1}$$

$$H_1: P_t = (p_{t-1}) - d^*$$

Requirement 1, above, leads to a type II error, $\beta=1-\pi$. We are then lead to solve the following inequalities simultaneously for n_t and x_t^* as follows. Let $B(x, n; p)$ represent the cumulative binomial probability of x or fewer successes in n trials. That is,

$$B(x, n; p) = \sum_{j=0}^x \binom{n}{j} p^j (1-p)^{n-j}$$

Then the inequalities of interest are:

$$B(x_t^*, n_t; p_t) \leq \alpha \quad (1)$$

$$B(x_t^*, n_t; p_t - d^*) \geq 1 - \beta \quad (2)$$

For p_{t-1} known, the null hypothesis is rejected if the current sample yields x_t^* or fewer reliable missiles. Since p_{t-1} is not known, however, (1) and (2) are solved after substituting p_{t-1} for p_t since we have no target value for it. We will account for this uncertainty by averaging the pair x_t^*, n_t with respect to the prior distribution for p_t . First, however, we shall introduce a sequential scheme to reduce the sample sizes required.

For practical reasons, the missiles are tested sequentially in time. Therefore, when a critical sample value is obtained, the sampling may be curtailed. That is, if x_t^*+1 successful tests or if $n_t - x_t^*+1$ failures are experienced before the sample is completed, then the test may be curtailed (terminated prematurely) without effecting the error distribution of the test. The curtailed sampling distribution is expressed as follows. Given p_t and x_t^* , the probability that $n_t=x$ when a curtailed sampling procedure is used is:

$$P[n_t=x | p_t] = \begin{cases} \binom{x-1}{n_t-x_t^*-1} (1-p_t)^{n_t-x_t^*} p_t^{x-(n_t-x_t^*)}, & n_t-x_t^* \leq x \leq x_t^* \\ \binom{x-1}{n_t-x_t^*-1} (1-p_t)^{n_t-x_t^*} p_t^{x-(n_t-x_t^*)} + \\ \binom{x-1}{x-x_t^*-1} (1-p_t)^{x-x_t^*-1} p_t^{x_t^*+1}, & x_t^* < x \leq n_t \end{cases} \quad (3)$$

In order to obtain $P[n_t=x]$, we compute the average with respect to the prior probability for p_t , given by $g(p_t | H)$. In the absence of information to the contrary, the conjugate prior in

the binomial case, is not only convenient, but also natural. This prior is the Beta distribution given by:

$$g(p|a,b,H) = B^{-1}(a,b)p^{a-1}(1-p)^{b-1}, \quad a \geq 1, \quad b \geq 1,$$

where,

$$B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b),$$

$\Gamma(x)$ is the gamma function [1, p. 255], and H refers to the experimental hypothesis relevant to our situation. The averaging process yields:

$$P[n_t=x] = \begin{cases} \binom{x-1}{n_t-x_t-1} B^{-1}(a,b) B(x-n_t+x_t+a, n_t-x_t+b) & \text{for } n_t-x_t \leq x \leq x_t \\ \binom{x-1}{n_t-x_t-1} B^{-1}(a,b) B(x-n_t+x_t+a, n_t-x_t+b) + & (4) \\ \binom{x-1}{x-x_t-1} B^{-1}(a,b) B(x_t+1+a, x-x_t-1+b) & \text{for } x_t < x \leq n_t \end{cases}$$

The expected sample size, $E[n_t]$, can be obtained by computing:

$$E[n_t] = \sum_{x=0}^n x P[n_t=x]$$

A full Bayesian treatment of the problem is developed as follows. Equations (1) and (2) are averaged with respect to the prior as shown below.

$$\int_0^1 B(x_t, n_t; p_t) g(p_t|H) dp_t \leq \alpha \quad (5)$$

$$\int_0^1 B(x_t, n_t; p_t - d) g(p_t|H) dp_t \geq 1 - \alpha \quad (6)$$

Integrals (5) and (6) may be re-expressed in closed form which allows them to be solved iteratively for x_t and n_t . These values are then used in equations (3) for computing the expected sample sizes. We point out that (5) is related to the predictive distribution which is used for model checking or informal hypothesis testing in the Bayesian context [2, p.385].

Generally, prior distributions on unknown parameters involve parameters of their own which, in turn, depend on the experimental conditions or hypotheses. In our example the parameters are 'a' and 'b'. The experimental hypotheses and specific parametric values for our situation are obtained and applied by using the following line of reasoning. Before the initial test, little or no a-priori information is available about p_1 so a flat prior distribution is assumed. The uniform prior corresponds to the parameter values $a=b=1$, and essentially assigns equal weights to all values of p_1 in the interval $(0,1)$. After the first test sample has been obtained, say x_1 and n_1 , the posterior distribution is a Beta distribution with parameters $a+x_1$ and $b+n_1-x_1$. The mode of the posterior may be used as an estimate for p_1 . This is given by $p_1=(a+x_1-1)/(a+b+n_1-2)$, and as noted previously, is the value against which the second sample is tested. The complete testing strategy is outlined below.

1. Before testing begins, the prior distribution is defined. This should be based on engineering knowledge and experience and developmental history. Since it is not usually possible to obtain that information from engineers, it is imperative to provide a reasonable alternative. For this we suggest using an initial sample, corresponding to time $t=0$. The implied prior for the initial sample is the uniform distribution of the Beta family.

2. The monitoring procedure begins with the first test sample and proceeds as follows. At time $t(=1,2,3,\dots)$ the prior distribution, $g_t(\cdot)$, is the posterior distribution from the test at time $t-1$, or $h_{t-1}(\cdot)$. The mode of the prior is the value for p_{t-1} in the null hypothesis against which the sample at time t is tested.

3. The sample size and critical value for the test is obtained from equations (1) and (2). If the sample results on an acceptance of the null hypothesis, then the sample values are used to update the prior, resulting in the posterior distribution. A new modal value for p is obtained which will be used in the test at time $t+1$, and a new sample size and critical value are obtained.

4. If the sample results in a rejection of the null hypothesis at time t , then the current prior is discarded, and the current sample is used to determine the prior for the following test of hypothesis.

The authors wish to acknowledge helpful discussions with Prof. George Box, Prof. Michael Woodroffe, and Dr. Daniel Willard.

REFERENCES

- [1] Abramowitz, Milton and Irene Stegun; 'Handbook of Mathematical Tables'; Dover Publications Inc., New York. (1964)
- [2] Box, George E. P.; 'Sampling and Bayes' Inference in Scientific Modeling and Robustness' JRSS(Series A), V 143, pps 383-430. (1980)

AN OPTIMAL SEQUENTIAL BERNOULLI
SELECTION PROCEDURE

Robert E. Bechhofer
School of Operations Research
and Industrial Engineering
College of Engineering
Cornell University
Ithaca, New York 14853

Research supported by
U.S. Army Research Office-Durham Contract DAAG-29-81-K-0168
at Cornell University.

Table of Contents

Abstract	1
1. Introduction	2
2. Statistical assumptions and notation	2
3. Bernoulli selection procedures	2
3.1 A single-stage procedure	2
3.2 Sequential procedures involving one-at-a-time sampling	4
3.2.1 Procedures $P_C = (R_C, S_C, T_C)$	5
3.2.2 The procedure $P^* = (R^*, S^*, T^*)$	7
4. Optimality properties of P^*	10
5. Performance of P^*	12
6. Concluding remarks	18
7. Acknowledgments	18
8. References	19

Abstract

This paper describes a new closed adaptive sequential procedure proposed by Bechhofer and Kulkarni [1982a] for selecting the Bernoulli population which has the largest success probability. The performance of this procedure is compared to that of the Sobel-Huyett [1957] single-stage procedure, and to a curtailed version of the single-stage procedure, all of which guarantee the same probability of a correct selection. Optimal properties of the Bechhofer-Kulkarni procedure are stated; quantitative assessments of important performance characteristics of the procedure are given. These demonstrate conclusively the superiority of the new procedure over that of the competing procedures. Relevant areas of application are described. Appropriate literature references are provided.

Key Words

Bernoulli selection problem, selection procedures, single-stage procedure, closed sampling procedures, one-at-a-time sampling procedures, adaptive sampling procedures, curtailed sampling procedures, vendor selection, clinical trials.

1. Introduction

The problem of devising statistical procedures for selecting the Bernoulli population which has the largest "success" probability has been the subject of intensive research by many investigators for more than twenty-five years. Interest in this problem stems from the fact that it arises in many areas of application of great practical importance: It arises, for example, in vendor selection when the purchaser seeks to identify the vendor with the largest fraction of conforming items. Similarly, in research and development, it arises when the scientist wishes to identify the process or system which has the largest probability of performing best. In clinical trials the medical researcher studies various treatment regimes with the intent of determining the one which has the largest probability of achieving a cure (or some other desirable effect) for the malady under investigation. Recently it has been shown that the problem of selecting the Bernoulli population with the largest success probability is closely related for quantal response curves to the problem of selecting the curve with the smallest q -quantile; this latter problem arises in certain military and medical settings. (See Tamhane [1983].)

The published literature on the Bernoulli selection problem and associated procedures is vast. The interested reader is referred to an article by Bechhofer and Kulkarni [1982a] for a recent survey of these papers. In that article the authors proposed closed adaptive sequential procedures for various Bernoulli selection goals. Unlike earlier procedures which had been proposed on ad hoc or heuristic grounds, these new procedures have certain important optimality properties and in addition have very desirable performance characteristics. It is the purpose of this present article to introduce the reader to the Bechhofer-Kulkarni procedure

for the particular goal of selecting the Bernoulli population which has the largest success probability, and to describe some of its properties. Appropriate literature references are given for those who wish to study the procedure in greater depth. It is perhaps of some interest to note that all of the articles concerning this new procedure have appeared within the past two years.

2. Statistical assumptions and notation

Let Π_i ($1 \leq i \leq k$) denote $k \geq 2$ Bernoulli populations with corresponding single-trial "success" probabilities p_i . Denote the ordered values of the p_i by $p_{[1]} \leq \dots \leq p_{[k]}$; the values of the p_i and of the $p_{[j]}$, and the pairing of the Π_i with the $p_{[j]}$ ($1 \leq i, j \leq k$) are assumed to be completely unknown. The goal of the experimenter is to select the population associated with $p_{[k]}$; when this population is selected, the experimenter is said to have made a correct selection (CS). For each of the examples cited in Section 1, it is meaningful to refer to the population associated with $p_{[k]}$ as the "best" population.

3. Bernoulli selection procedures

3.1 A single-stage procedure

Sobel and Huyett [1957] proposed a single-stage procedure for selecting the best Bernoulli population. Their procedure which was developed while the authors were employed at the Bell Telephone Laboratories was motivated by industrial applications. This single-stage procedure $P_{SS} = (R_{SS}, T_{SS})$ has a sampling rule (R_{SS}) and a terminal decision rule (T_{SS}) which are given below.

SINGLE-STAGE PROCEDURE (P_{SS}) FOR SELECTING THE POPULATION

ASSOCIATED WITH $P[k]$.

Sampling rule (R_{SS}): Take exactly n independent observations from every population. (3.1)

Terminal decision rule (T_{SS}): Let x_i denote the number of "successes" in the n observations from Π_i , and let $x_{[1]} \leq x_{[2]} \leq \dots \leq x_{[k]}$ denote the ordered values of the x_i ($1 \leq i \leq k$). Select the population that yielded $x_{[k]}$ as the one associated with $P[k]$, randomizing among all populations that have x -values equal to $x_{[k]}$. (3.2)

We now give two examples of P_{SS} . In these examples we denote a success (failure) from Π_i by S_i (F_i) ($1 \leq i \leq k$).

Example 1: ($k = 3, n = 3$)

<u>Π_1</u>	<u>Π_2</u>	<u>Π_3</u>
F_1	S_2	S_3
S_1	S_2	F_3
S_1	S_2	S_3

Here $x_{[2]} = 2 < x_{[3]} = 3$, which was yielded by Π_2 . Hence, select Π_2 as the population associated with $P_{[3]}$.

Example 2: ($k = 3, n = 3$)

π_1	π_2	π_3
S_1	F_2	S_3
S_1	S_2	S_3
S_1	S_2	S_3

Here $x_{[1]} = 2 < x_{[2]} = x_{[3]} = 3$ which was yielded by π_1 and π_3 . Hence, select (π_1, π_3) with probability $(\frac{1}{2}, \frac{1}{2})$ as the population associated with $p_{[3]}$.

3.2 Sequential procedures involving one-at-a-time sampling

Throughout we limit consideration to the class of sampling rules (R) which take no more than n observations from any one of the k populations; the single-stage procedure is clearly in this class. The choice of $n \geq 1$ is arbitrary and can be arrived at using economic considerations.

We shall describe sequential procedures in which observations are taken one-at-a-time (instead of in a single-stage) and show the gains that can be achieved by employing them. We denote a success (failure) from π_i at stage m by S_i^m (F_i^m) ($1 \leq i \leq k, 1 \leq m \leq kn$). Let $n_{i,m}$ denote the total number of observations taken from π_i through stage m , and let $z_{i,m}$ denote the total number of successes yielded by π_i through stage m ($1 \leq i \leq k; 1 \leq m \leq kn$).

In Section 3.2.1 we describe a sequential procedure $P_C = (R_C, S_C, T_C)$ which uses arbitrary one-at-a-time sampling rules in conjunction with an obvious stopping rule employing what we term weak curtailment (along with

an associated terminal decision rule). We show by examples how P_C can achieve a decrease in the total number of observations to termination relative to the kn observations required by the corresponding single-stage procedure. In Section 3.2.2 we describe our new sequential procedure $P^* = (R^*, S^*, T^*)$ which uses an optimal one-at-a-time sampling rule in conjunction with a stopping rule employing what we term strong curtailment (along with an associated terminal decision rule). Examples are given to show how P^* operates, and the savings that can be achieved by using it. By weak curtailment we mean that a strict inequality ($>$) holds in (3.4) below, while by strong curtailment we mean that a weak inequality ($\geq$) holds in (3.6) below.

In Section 4 we give some of the optimality properties of P^* , and point out the ways in which P^* is superior to P_C . Section 5 contains some typical results concerning the performance of P^* . We make some concluding remarks in Section 6.

3.2.1 Procedures $P_C = (R_C, S_C, T_C)$

We now describe the procedures (P_C) in this class.

PROCEDURES (P_C) FOR SELECTING THE POPULATION ASSOCIATED WITH $p_{[k]}$.

Sampling rule (R_C) : At stage m ($0 \leq m \leq kn$), take the next observation from an arbitrary one of the k populations. (3.3)

Stopping rule (S_C) : Stop sampling at the first stage m at which there exists at least one population π_i satisfying (3.4)

$$z_{i,m} > z_{j,m} + n - n_{j,m} \text{ for all } j \neq i \text{ } (1 \leq i, j \leq k).$$

Terminal decision rule (T_C): If $r \geq 1$ populations, say

$\Pi_{i_1}, \dots, \Pi_{i_r}$, simultaneously satisfy (3.4), then select one of them at random as associated with $p_{[k]}$.

The stopping sequences given in the following examples illustrate how

$P_C = (R_C, S_C, T_C)$ operates.

Example 3: ($k = 3, n = 3$)

Π_1	Π_2	Π_3
F_1^3	S_2^4	S_3^1
	S_2^5	F_3^2
	S_2^6	

In this example we have assumed that the first six outcomes of Example 1 were obtained in the order indicated by the superscripts. Clearly one can stop sampling after having obtained S_2^6 , and select Π_2 as the population associated with $p_{[3]}$. Here Π_2 has "beaten" Π_1 and Π_3 , and this result will not change no matter what the outcomes of the remaining three observations.

Example 4: ($k = 3, n = 3$)

Π_1	Π_2	Π_3
S_1^2	F_2^1	S_3^5
S_1^3		S_3^6
S_1^4		S_3^7

In this example we have assumed that the first seven outcomes of Example 2 were obtained in the order indicated by the superscripts. Thus one can stop sampling after having obtained S_3^7 , and select (π_1, π_3) with probability $(\frac{1}{2}, \frac{1}{2})$ as the population associated with $P_{[3]}$.

Remark 3.1: We see that P_C arrives at the same terminal decision as does P_{SS} . Therefore, it achieves the same probability of a correct selection as does P_{SS} . Moreover, it usually accomplishes this with a smaller total number of observations to termination than the kn observations of the corresponding P_{SS} .

3.2.2 The procedure $P^* = (R^*, S^*, T^*)$

Our procedure P^* which uses an optimal sampling rule (R^*) in conjunction with the stopping rule (S^*) and the terminal decision rule (T^*) is described below.

PROCEDURE (P^*) FOR SELECTING THE POPULATION ASSOCIATED WITH $P_{[k]}$.

Sampling rule (R^*) : At stage m ($0 \leq m \leq kn-1$), take the next observation from the population which has the smallest number of failures among all π_i for which $n_{i,m} < n$ ($1 \leq i \leq k$). If there is a tie among such equal-number-of-failure populations, take the next observation from (3.5) that one of them that has the largest number of successes. If there is a further tie among such equal-number-of-success populations, select one of them at random and take the next observation from it.

Stopping rule (S^*): Stop sampling at the first stage m at

which there exists at least one population π_i

satisfying

(3.6)

$$z_{i,m} \geq z_{j,m} + (n - n_{j,m}) \text{ for all } j \neq i \quad (1 \leq i, j \leq k).$$

Terminal decision rule (T^*): If $r \geq 1$ populations, say

$\pi_{i_1}, \dots, \pi_{i_r}$, simultaneously satisfy (3.6), then

(3.7)

select one of them at random as associated with $p_{[k]}$.

We now give two stopping sequences to illustrate how

$P^* = (R^*, S^*, T^*)$ operates.

Example 5: ($k = 3, n = 3$)

π_1	π_2	π_3
F_1^3	S_2^4	S_3^1
	S_2^5	F_3^2

In this example we have applied P^* to the first five outcomes of Example 3. We see that at stage 5, π_2 satisfies (3.6). Hence, select π_2 as associated with $p_{[3]}$. Here neither π_1 nor π_3 can do better than tie π_2 no matter what the outcomes of the remaining four observations.

Note: We point out that (3.5) is a well-defined sampling rule which dictates the population or populations from which the next observation must be taken. Thus, for example, if in Example 5 the outcome of the second observation from π_3 were a success (S_3^2) instead of a failure (F_3^2), then the third observation must be taken from π_3 .

Example 6: ($k = 3, n = 3$)

$$\begin{array}{ccc} \underline{\pi_1} & \underline{\pi_2} & \underline{\pi_3} \\ S_1^2 & F_2^1 & \\ S_1^3 & & \\ S_1^4 & & \end{array}$$

In this example we have applied p^* to the first four outcomes of Example 4. We see that at stage 4, π_1 satisfies (3.6). Hence, select π_1 as associated with $p_{[3]}$. Here π_1 has "beaten" π_2 , and π_3 cannot do better than tie π_1 , no matter what the outcomes of the remaining five observations.

We now point out an important property shared by P_{SS} , P_C and p^* (along with many other competing procedures). This property is summarized in Theorem 3.1, below. In the theorem it is assumed that if two or more populations have a common p -value equal to $p_{[k]}$, then these tied populations are tagged in such a way that their ordering is unique, i.e., one is associated with $p_{[k]}$, a second with $p_{[k-1]}$, etc.

Theorem 3.1: $P\{CS|(R_{SS}, T_{SS})\} \equiv P\{CS|(R^*, S^*, T^*)\}$ uniformly in $(p_1, p_2, \dots, p_k)$.

This fundamental result was first proved (in more generality and under very reasonable assumptions) in Kulkarni [1981], and reported in Bechhofer and Kulkarni [1982a]. More recently Jennison [1983] proved a much more general result.

Note: We have already pointed out in Remark 3.1 that $P\{CS|(R_{SS}, T_{SS})\} \equiv P\{CS|(R_C, S_C, T_C)\}$ uniformly in $(p_1, p_2, \dots, p_k)$.

In order to have a rational basis for making distinctions between these procedures, and in particular for deciding which one is "best" in some reasonable sense, it is necessary to study other important performance characteristics of the procedures. Two such performance characteristics are $E\{N_{(i)}\}$ ($1 \leq i \leq k$) and $E\{N\}$; here $N_{(i)}$ denotes the total number of observations taken from the population associated with $p_{[i]}$ ($1 \leq i \leq k$) and $N = \sum_{i=1}^k N_{(i)}$ denotes the total number of observations taken from all k populations, when a procedure terminates sampling. In Section 4 we cite several optimality properties of P^* , stated in terms of $E\{N_{(i)}\}$ ($1 \leq i \leq k$) and $E\{N\}$. In Section 5 we give some typical results of studies made of the performance of $E\{N\}$.

4. Optimality properties of P^*

The theorems cited below concerning the optimality of P^* are proved in their present generality (along with others) in Kulkarni and Jennison [1983], and are reported on in Bechhofer and Frisardi [1983]. Earlier, more restricted versions were proved by Kulkarni [1981]. Further optimality results are contained in Jennison and Kulkarni [1984].

In this section, R refers (as before) to an arbitrary sampling rule which takes no more than n observations from any one of the $k \geq 2$ populations, and which is used in conjunction with the stopping rule S^* and the terminal decision rule T^* of (3.6) and (3.7), respectively. For $k = 2$ let $\bar{R}^*$ denote the conjugate sampling rule in which $n_{i,m} - z_{i,m}$ and $z_{i,m}$ of (3.6) are replaced by $z_{i,m}$ and $n_{i,m} - z_{i,m}$, respectively. We now state several theorems concerning the optimality of R^* and $\bar{R}^*$.

Theorem 4.1: For $k = 2$, a necessary and sufficient condition that

$P^* = (R^*, S^*, T^*)$ minimize $E\{N|(p_1, p_2)\}$ among all procedures (R, S^*, T^*)

is $p_1 + p_2 \geq 1$. For $k = 2$, a necessary and sufficient condition that

$\bar{P}^* = (\bar{R}^*, S^*, T^*)$ minimize $E\{N|(p_1, p_2)\}$ among all procedures (R, S^*, T^*)

is $p_1 + p_2 \leq 1$.

Theorem 4.2: For $k = 2$, a necessary and sufficient condition that

$P^* = (R^*, S^*, T^*)$ minimize $E\{N_{(1)}|(p_1, p_2)\}$ among all procedures

(R, S^*, T^*) is

$$p_{[2]} \geq \frac{3 - p_{[1]} - \sqrt{(3 - p_{[1]})^2 - 4}}{2}. \quad (4.1)$$

Theorem 4.3: For $k \geq 3$, a sufficient condition that $P^* = (R^*, S^*, T^*)$

minimize $E\{N|(p_1, p_2, \dots, p_k)\}$ among all procedures (R, S^*, T^*) is

$$p_{[1]} + \sum_{i=2}^k p_{[i]} / (k-1) \geq 1. \quad (4.2)$$

Theorem 4.4: For $k \geq 3$, a sufficient condition that $P^* = (R^*, S^*, T^*)$

minimize $\sum_{i=1}^s E\{N_{(i)}|(p_1, p_2, \dots, p_k)\}$ for all s ($1 \leq s \leq k$) among all procedures (R, S^*, T^*) is $p_{[1]} + p_{[2]} \geq 1$.

Remark 4.1: It can be shown from (4.1) that P^* minimizes $E\{N_{(1)}|(p_1, p_2)\}$ among all procedures (R, S^*, T^*) over approximately 81.55 percent of the (p_1, p_2) -parameter space.

Remark 4.2: In the context of clinical trials it is desirable to minimize the expected number of observations taken from the inferior populations, i.e., those with small p-values. Hence, the relevance of Theorems 4.2 and 4.4.

The foregoing theorems summarize some of the most important optimality properties of P^* , and show why, in particular, it is superior to P_{SS} and P_C . Not only is P^* superior to P_{SS} and P_C but also any procedure which uses R in conjunction with S^* is superior to a corresponding procedure which uses the same R in conjunction with S_C .

Remark 4.3: The following additional properties of P^* have been shown to hold:

- a) $n \leq N \leq kn-1$ for all $(p_1, p_2, \dots, p_k)$
- b) $P\{N = n \mid (p_1, p_2, \dots, p_k)\} \rightarrow 1$ for $p_{[1]} \rightarrow 1$,
 $P\{N = kn-1 \mid (p_1, p_2, \dots, p_k)\} \rightarrow 1$ for $p_{[k]} \rightarrow 0$.

As a consequence of a) we have

- c) $n \leq E\{N\} \leq kn-1$ for all $(p_1, p_2, \dots, p_k)$
- d) $\frac{1}{k} \leq \frac{E\{N\}}{kn} \leq \frac{kn-1}{kn}$ for all $(p_1, p_2, \dots, p_k)$. The ratio $E\{N\}/kn$ can be thought of as a measure of relative efficiency, small values of the ratio favoring P^* .

5. Performance of P^*

Extensive studies of the behavior of P^* have been carried out in order to obtain numerical assessments of its performance--in particular to study the distribution of $N_{(i)}$, and $E\{N_{(i)}\}$ ($1 \leq i \leq k$), the distribution of N , and $E\{N\}$, as well as the achieved $P\{CS\}$. Bechhofer and

Kulkarni [1982b] provides many tables of these quantities for $k = 2$ and 3 with selected n and $(p_{[1]}, p_{[2]}, \dots, p_{[k]})$; all of the results given in the tables are exact, having been calculated using recursion formulae.

Bechhofer and Frisardi [1983] provides a large number of analogous tables containing very precise estimates of such quantities (and others) for $k = 3, 4$ and 5 with selected n and $(p_{[1]}, p_{[2]}, \dots, p_{[k]})$; these were obtained using Monte Carlo (MC) simulation since the cost of calculating exact results would have been prohibitive.

Three typical tables taken from the aforementioned articles are reproduced here. Table 5.1 shows for $k = 3, n = 7$ how the distribution of $N_{(i)}$ ($1 \leq i \leq 3$) and hence $E\{N_{(i)}\}$ ($1 \leq i \leq 3$) and $E\{N\}$ change as the differences between the $p_{[i]}$ ($1 \leq i \leq 3$) become larger; in each case the p_i ($1 \leq i \leq 3$) are equally-spaced around $p_{[2]} = 0.6$, the spacing increasing from 0.1 to 0.4 . We note the dramatic decrease in $E\{N_{(1)}\}$ and $E\{N_{(2)}\}$, and also the large decrease in $E\{N\}$ as the spacing increases. The $E\{N_{(i)}\}$ ($1 \leq i \leq 3$) values of $3.47, 4.28$ and 5.46 for the p -vector $(0.5, 0.6, 0.7)$, and the values $0.62, 1.22$ and 6.58 for the p -vector $(0.2, 0.6, 1.0)$ are to be compared with the $n = 7$ observations per population required by the corresponding single-stage procedure; the corresponding $E\{N\}$ values of 13.21 and 8.42 are to be compared with $kn = 21$ for the single-stage procedure. As a consequence of Theorem 4.3 we note that P^* is optimal for both of the p -vectors in Table 5.1 since $p_{[1]} + (p_{[2]} + p_{[3]})/2 \geq 1$.

Table 5.2 shows for $k = 5, n = 50$, how $E\{N_{(i)}\}$ ($1 \leq i \leq 5$) decreases as $p_{[5]}$ of the p -vector $(p_{[1]}, p_{[2]}, p_{[3]}, p_{[4]}, p_{[5]})$ increases ($p_{[5]} = 0.45(0.10)0.95$) while the differences $p_{[i]} - p_{[i-1]}$ ($2 \leq i \leq 5$)

TABLE 5.1^{1/}

Exact distribution of $N_{(i)}$, and $E\{N_{(i)}\}$ ($i = 1, 2, 3$) and $E\{N\}$ for P^* when $k = 3$, $n = 7$ and $(p_{[1]}, p_{[2]}, p_{[3]}) = (0.5, 0.6, 0.7)$ and $(0.2, 0.6, 1.0)$

a	$p_{[1]}=0.5, p_{[2]}=0.6, p_{[3]}=0.7$			$p_{[1]}=0.2, p_{[2]}=0.6, p_{[3]}=1.0$		
	$P\{N_{(1)}=a\}$	$P\{N_{(2)}=a\}$	$P\{N_{(3)}=a\}$	$P\{N_{(1)}=a\}$	$P\{N_{(2)}=a\}$	$P\{N_{(3)}=a\}$
0	0.054	0.045	0.018	0.505	0.500	0.014
1	0.180	0.123	0.054	0.396	0.200	0.000
2	0.173	0.123	0.064	0.079	0.120	0.000
3	0.149	0.117	0.070	0.016	0.072	0.000
4	0.118	0.103	0.071	0.003	0.043	0.000
5	0.087	0.085	0.066	0.001	0.026	0.000
6	0.081	0.109	0.146	0.000	0.016	0.324
7	0.158	0.294	0.512	0.000	0.023	0.662
$E\{N_{(1)}\}$	3.47	---	---	0.62	---	---
$E\{N_{(2)}\}$	---	4.28	---	---	1.22	---
$E\{N_{(3)}\}$	---	---	5.46	---	---	6.58
$E\{N\}$	$(3.47 + 4.28 + 5.46) = 13.21$			$(0.62 + 1.22 + 6.58) = 8.42$		

^{1/} Abstracted from Table 4.13 of Bechhofer and Kulkarni [1982b].

Note: The corresponding single-stage procedure (which guarantees exactly the same probability of a correct selection as P^*) requires seven observations from each of the three populations.

TABLE 5.2^{1/}

Monte Carlo estimates of $E\{N_{(i)}\}$ ($1 \leq i \leq 5$) for P^*
 when $k = 5$, $n = 50$ for selected $(p_{[1]}, p_{[2]}, \dots, p_{[5]})$

$(p_{[1]}, p_{[2]}, \dots, p_{[5]})$	Monte Carlo estimate of				
	$E\{N_{(1)}\}$	$E\{N_{(2)}\}$	$E\{N_{(3)}\}$	$E\{N_{(4)}\}$	$E\{N_{(5)}\}$
(0.05, 0.15, 0.25, 0.35, 0.45)	28.56	31.94	35.99	41.96	49.13
(0.15, 0.25, 0.35, 0.45, 0.55)	25.74	29.33	34.03	39.96	49.10
(0.25, 0.35, 0.45, 0.55, 0.65)	22.33	25.90	30.61	37.57	48.86
(0.35, 0.45, 0.55, 0.65, 0.75)	18.99	22.16	27.02	34.91	48.79
(0.45, 0.55, 0.65, 0.75, 0.85)	13.26	15.86	20.40	29.50	48.75
(0.55, 0.65, 0.75, 0.85, 0.95)	5.28	6.55	9.39	17.17	48.99

^{1/} Taken from Table II of Bechhofer and Frisardi [1983] with results for $E\{N_{(5)}\}$ added.

Note: The corresponding single-stage procedure (which guarantees exactly the same probability of a correct selection as P^*) requires fifty observations from each of the five populations.

TABLE 5.3^{1/}

Monte Carlo estimates of $E\{N\}$ for P^* when $k = 5$,
 $n = 10, 30$ and 50 for selected $(p_{[1]}, p_{[2]}, \dots, p_{[5]})$

$(p_{[1]}, p_{[2]}, \dots, p_{[5]})$	Monte Carlo estimate of $E\{N\}$		
	$kn = 50$	$kn = 150$	$kn = 250$
$(0.05, 0.15, 0.25, 0.35, 0.45)$	34.48	110.28	187.55
$(0.15, 0.25, 0.35, 0.45, 0.55)$	31.54	104.81	178.16
$(0.25, 0.35, 0.45, 0.55, 0.65)$	29.57	98.54	165.27
$(0.35, 0.45, 0.55, 0.65, 0.75)$	26.15	89.75	151.87
$(0.45, 0.55, 0.65, 0.75, 0.85)$	22.09	75.04	127.76
$(0.55, 0.65, 0.75, 0.85, 0.95)$	17.46	54.40	87.38

^{1/}This is Table VI of Bechhofer and Frisardi [1983].

Note: The corresponding single-stage procedures (which guarantees exactly the same probability of a correct selection as P^*) require exactly n observations from each population.

remain equal to 0.1. While p^* is only known to be optimal here for the p -vectors with $p_{[5]} = 0.85$ and 0.95 we see that the $E\{N_{(i)}\}$ ($1 \leq i \leq 4$) decrease dramatically for increasing $p_{[5]}$, always being substantially less than the $n = 50$ required from each population by the single-stage procedure which guarantees the same $P\{CS\}$.

Finally, Table 5.3 shows for $k = 5$ and $n = 10, 30, 50$ and the same p -vectors as used in Table 5.2, how $E\{N\}$ decreases as $p_{[5]}$ increases; see Remark 4.3 for an explanation of this phenomenon. Here the $E\{N\}$ -values in any column are to be compared with the kn -value required by the single-stage procedure which guarantees the same $P\{CS\}$; thus, for example, each entry in the third column is to be compared to $kn = 250$.

We see from Tables 5.1 and 5.2 that p^* tends to sample far less frequently on the average from the inferior populations than it does from the superior populations; this is highly desirable in clinical trials. Table 5.3 shows that $E\{N\}$ decreases as the $p_{[i]}$ ($1 \leq i \leq 5$) increase; this is highly desirable in vendor selection where most of the $p_{[i]}$ ($1 \leq i \leq 5$) tend to be large. The results cited in these tables are typical of those given in the tables of Bechhofer-Kulkarni [1982b] and Bechhofer-Frisardi [1983].

Remark 5.1: General methods for estimating and bounding $E\{N_{(i)}\}$ ($1 \leq i \leq k$) and $E\{N\}$ for p^* are given in Jennison [1984]; these improve on earlier results given in Bechhofer and Kulkarni [1982b].

6. Concluding remarks

We have demonstrated conclusively that P^* has highly desirable performance characteristics. Very substantial savings in $E\{N\}$ can be realized if P^* is used in place of the Sobel-Huyett single-stage procedure with both achieving the same $P\{CS\}$; these savings increase as the p_i -values ($1 \leq i \leq k$) increase. In addition P^* samples from the inferior populations far less than from the superior ones thus making it particularly attractive for clinical trials. Finally, we note that from a practical point of view, P^* is very easy to carry out, and no special tables are needed for its implementation.

7. Acknowledgments

This research was supported by the U.S. Army Research Office-Durham under Contract DAAG29-81-K-0168 at Cornell University. The writer acknowledges with thanks the helpful comments of Dr. Radhika V. Kulkarni and Professor Ajit C. Tamhane.

8. References

- Bechhofer, R.E. and Frisardi, T. (1983): A Monte Carlo study of the performance of a closed adaptive sequential procedure for selecting the best Bernoulli population. Journal of Statistical Computation and Simulation, 18, 179-213.
- Bechhofer, R.E. and Kulkarni, R.V. (1982a). Closed adaptive sequential procedures for selecting the best of $k \geq 2$ Bernoulli populations. Proceedings of the Third Purdue Symposium on Statistical Decision Theory and Related Topics (ed. by S.S. Gupta and J. Berger), New York, Academic Press, I, 61-108.
- Bechhofer, R.E. and Kulkarni, R.V. (1982b). On the performance characteristics of a closed adaptive sequential procedure for selecting the best Bernoulli population. Sequential Analysis: Communications in Statistics (C), 1(4), 315-354.
- Jennison, C. (1983). Equal probability of correct selection for Bernoulli selection procedures. Communications in Statistics--Theory and Methods, A12, 24, 2887-2896.
- Jennison, C. (1984). On the expected sample size for the Bechhofer-Kulkarni Bernoulli selection procedure. Sequential Analysis: Communications in Statistics (C), 3(1).
- Jennison, C. and Kulkarni, R.V. (1984). Optimal procedures for selecting the best s out of k populations. To appear in Essays in Honor of Robert E. Bechhofer, Marcel-Dekker.
- Kulkarni, R.V. (1981). Closed adaptive sequential procedures for selecting the best of $k \geq 2$ Bernoulli populations. Ph.D. dissertation, Cornell University, Ithaca, New York.
- Kulkarni, R.V. and Jennison, C. (1983). On the optimal properties of the Bechhofer-Kulkarni Bernoulli sequential selection procedure. Submitted for publication.
- Sobel, M. and Huyett, M. (1957). Selecting the best one of several binomial populations. Bell System Technical Journal, 36, 537-576.
- Tamhane, A.C. (1983). A survey of literature on estimation methods for quantal response curves with a view toward applying them to the problem of selecting the curve with the smallest q -quantile (ED_{100q}). Preliminary Report, Technical Report No. 614, School of Operations Research and Industrial Engineering, Cornell University.

A New Method of Evaluating Normal and t-Tail Areas

Andrew P. Soms*

Department of Mathematical Sciences
University of Wisconsin-Milwaukee, Milwaukee, WI 53201
Mathematics Research Center, 610 Walnut Street
Madison, WI 53705

Key Words and Phrases: absolute error, Bonferroni percentiles, normal distribution, relative error, t-distribution.

Abstract

The bounds of Boyd (1959) and Soms (1980a, 1980b) for the tail areas of the normal and t-distributions are used to obtain a new method of evaluating the tail areas. The absolute and relative errors and numerical examples are given.

*This research was supported by the Mathematics Research Center, University of Wisconsin-Madison, under Contract No. DAAG29-80-C-0041 and the University of Wisconsin-Milwaukee.

1. The Method

We begin by introducing notation and stating the main results of Boyd (1959) and Soms (1980a, 1980b). Boyd (1959) showed that if

$$\phi(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2), \quad \bar{F}(x) = \int_x^{\infty} \phi(t) dt,$$

and $R_x = \bar{F}(x)/\phi(x)$, $x > 0$, then

$$p(x, \gamma_{\min}) < R_x < p(x, \gamma_{\max}),$$

where $p(x, \gamma) = (\gamma + 1)/[(x^2 + (2/\pi)(\gamma + 1)^2)^{\frac{1}{2}} + \gamma x]$, $\gamma_{\max} = 2/(\pi - 2)$,

$\gamma_{\min} = \pi - 1$, and the bounds are the best possible in the class

$\{p(x, \gamma), \gamma > -1\}$. This is also discussed in Johnson and Kotz (1970, Ch. 33).

Soms (1980a, 1980b) extended the above results and showed that if for arbitrary real $k > 0$ and $x > 0$,

$$f_k(t) = c_k (1+t^2/k)^{-(k+1)/2}, \quad c_k = \frac{\Gamma((k+1)/2)}{\Gamma(k/2)(\pi k)^{1/2}},$$

$$\bar{F}_k(x) = 1 - F_k(x) = \int_x^{\infty} f_k(t) dt,$$

$$R_k(x) = \bar{F}_k(x)/[(1+x^2/k)f_k(x)],$$

for $k > 2$, $\gamma_{\max} = 4c_k^2/(1-4c_k^2)$ and $\gamma_{\min} = \frac{k}{2(k+2)c_k^2} - 1$, and for

$k < 2$, $\gamma_{\min}$ and $\gamma_{\max}$ are interchanged,

and

$$p(x, \gamma) = \frac{1+\gamma}{2 \frac{2}{2} \frac{2}{2} \frac{2}{2} \frac{1}{2}},$$

then

$$p(x, \gamma_{\min}) < R_k(x) < p(x, \gamma_{\max}) ,$$

or equivalently,

$$(1 + \frac{x^2}{k}) f_k(x) p(x, \gamma_{\min}) < \bar{F}_k(x) < (1 + \frac{x^2}{k}) f_k(x) p(x, \gamma_{\max}) ,$$

and the bounds again are best in the same sense as for the normal.

It was also shown there that if $k = 2$, $\gamma_{\max} = \gamma_{\min} = \gamma_2$ and $R_k(x) = p(x, \gamma_2)$.

The numerical properties of these bounds are discussed in the above references. The important fact to be noted here is that the bounds control both absolute and relative error. Using the bounds as a starting point we now develop a simple method of evaluating normal and t-tail areas that controls both absolute and relative error, as opposed to the usual methods, which generally only control absolute error.

We consider estimates of the tail area of the form

$$(\frac{a+bx}{c+dx}) p(x, \gamma_{\min}) \phi(x) + (1 - \frac{a+bx}{c+dx}) p(x, \gamma_{\max}) \phi(x) \quad (1.1)$$

for the tail area of the normal and

$$(\frac{a+bx}{c+dx}) p(x, \gamma_{\min}) f_k(x) + (1 - \frac{a+bx}{c+dx}) p(x, \gamma_{\max}) f_k(x) \quad (1.2)$$

for the tail area of the t. We want the estimates to lie between the upper and lower bounds for the tail area and be strictly decreasing functions of x and therefore impose the added restrictions that

$$bc > ad$$

and

$$0 \leq \frac{a+bx}{c+dx} \leq 1, \quad \text{all } x \geq 0.$$

Since $f(0) = \frac{a}{c}$, we may, without loss of generality, assume that $c = 1$ and so our weight functions f are of the type

$$f(x) = \frac{a+bx}{1+dx}, \quad (1.3)$$

where $0 \leq a \leq 1$, $d > 0$, $bc > ad$, and $\frac{b}{d} \leq 1$. We then seek that particular choice of f which minimizes the absolute error. A direct computer search led to

$$f(x) = \frac{.71x}{1+.71x} \quad (1.4)$$

for the normal and

$$f(x) = \frac{b_k x}{1+b_k x}, \quad (1.5)$$

$$b_k = .70 + 1.82/k - .2/k^2, \quad (1.6)$$

for the t , where, as noted before, k is the degrees of freedom. (1.6)

was obtained by finding the optimal constants for $k = 25, 10, 5, 3, 1.5, 1, .5$ and fitting a regression line to them. However, in the interests of simplicity, for $k \leq 2$, we did not interchange $\gamma_{\min}$ and $\gamma_{\max}$ and so (1.5) and (1.6) are understood to apply for all k with $\gamma_{\min}$ and $\gamma_{\max}$ defined as for $k > 2$. Numerical evidence indicates that, at least for $k = 1$, the above optimal

estimate is still a decreasing function of x .

The maximum absolute and relative errors of the optimal estimates are remarkably constant over the range $1 \leq k \leq \infty$ and hence we only give the normal figures. For (1.4), the maximum absolute error is $.66 \times 10^{-4}$ and the maximum relative error is $.97 \times 10^{-3}$. We emphasize once more, that, unlike the usual methods, which generally control only absolute error, the above controls both absolute and relative error and hence can be used to calculate ordinary and Bonferroni descriptive levels and ordinary and Bonferroni percentiles.

As a check, we calculated the standard textbook table of the normal, given, e.g., in Brown and Hollander (1977) and found at most a difference of 1 in the fourth decimal place. We also compared the small normal percentiles given in Abramowitz and Stegun (1965, p. 977) to the ones obtained from (1.4) and after rounding both to three decimal places found that there was at most a difference of 1 in the third decimal place. Similar results apply to the t .

2. Concluding Remarks

We have given a method of calculating normal and t -tail areas which controls both absolute and relative errors. The listings of the short FORTRAN programs are available on request from the author. Preliminary results indicate that it is possible to improve on the accuracy of the approximations here described at a modest increase in complexity and these results will be reported shortly.

Bibliography

- Abramowitz, M. and Stegun, I. A. (1965), Handbook of Mathematical Functions, National Bureau of Standards, Washington, D.C.
- Brown, W. and Hollander, M. (1977), Statistics A Biomedical Introduction, John Wiley & Sons, New York.
- Johnson, N. L., and Kotz, S. (1970), Continuous Univariate Distributions-2, Houghton Mifflin Co., Boston.
- Boyd, A. V. (1959), "Inequalities for Mills' Ratio," Reports of Statistical Applications Research, Japanese Union of Scientists and Engineers, 6, 44-6.
- Soms, A. P. (1980a), "Rational Bounds for the t-Tail Area," Journal of the American Statistical Association, 75, 438-40.
- Soms, A. P. (1980b), "A Note on an Extension of Rational Bounds for the t-Tail Area to Arbitrary Degrees of Freedom," Technical Summary Report No. 2106, Mathematics Research Center, University of Wisconsin-Madison and to appear in Communications in Statistics.

THE DESIGN OF A QUANTAL RESPONSE EXPERIMENT:
AN EMPIRICAL APPROACH

Refik Soyer

The George Washington University
School of Engineering and Applied Science
Institute for Reliability and Risk Analysis

ABSTRACT
of
GWU/IRRA/TR-83/2
18 April 1983

An important issue in quantal response estimation problems considered by Mazzuchi and Singpurwalla (1982) is the design of the experiment. The objective there was to estimate the probability of response for a given stimulus, but due to the expense of the items, testing has to be kept to a minimum. As a continuation of the work by Mazzuchi and Singpurwalla (1982), we address this problem and present a criterion for comparing several designs that are of interest.

Research Supported by
Contract N00014-77-C-0263
Project Nr-042-372
Office of Naval Research
and
Grant DAAG-29-80-C-0067
Army Research Office

THE GEORGE WASHINGTON UNIVERSITY
School of Engineering and Applied Science
Institute for Reliability and Risk Analysis

THE DESIGN OF A QUANTAL RESPONSE EXPERIMENT:
AN EMPIRICAL APPROACH

Refik Soyer

1. INTRODUCTION

The U.S. Army Kinetic Energy Penetrator problem has been described by Mazzuchi and Singpurwalla (1982), henceforth MS. Their objective was to estimate the relationship between the striking velocity (the stimulus) and the probability of penetration of a projectile. This is a quantal response experiment in which the goal is to estimate the probability of response for a given stimulus.

The strategy used to test the effectiveness of the penetrator is to fix an angle of fire and then to fire the penetrator at different striking velocities. After each firing, the outcome, success or failure, is recorded.

The equipment used in testing is expensive, and thus testing is kept to a minimum. Typically, an experimenter is allowed a fixed number of tests. That is, a fixed number of copies of the penetrator can be tested at different striking velocities. Therefore, designing the experiment in an optimal way is an important issue. In a quantal response

problem, the investigator is also interested in estimating the striking velocity (stimulus), say V_α , at which the probability of penetration (response) is α . Thus, the experiment should be designed in a way that will provide the investigator with a "good" estimate of the V_α , for a specified amount of testing.

In this report, we attempt to present an approach that may be helpful in designing an experiment which addresses the objectives mentioned above. Due to the nature of the penetrator problem, interest generally centers around $V_{.05}$ and $V_{.95}$, stimuli at which the probabilities of response are 0.05 and 0.95, respectively. In our analysis, we will focus attention on the former.

2. AN OUTLINE OF THE APPROACH

Suppose that the experimenter is allowed to test k copies of the penetrator at k distinct levels of the stimuli. Our goal is to select the k distinct levels of stimulus in a way that will provide us a "good" estimate of $V_{.05}$.

To estimate $V_{.05}$, we first estimate the response curve based on the k distinct firings. The approach discussed in MS is adopted.

Let $V_1 < V_2 < \dots < V_k$ be k distinct levels of the stimulus. Since our aim is to select these k distinct levels in an "optimal" way, different designs have to be considered in the analysis. Because actual testing under the various designs is not practically feasible, our analysis is based on a simulation.

2.1 Simulation of the Responses

The outcome of a test at V_i is described by a binary variable X_i , $i = 1, 2, \dots, k$, where $X_i = 1$ if the target is defeated and $X_i = 0$ otherwise. To simulate the outcome X_i of a test at stimulus V_i , we assume that we know the "true" probability of response at V_i , $i = 1, 2, \dots, k$.

Let $V_1 < V_2 < \dots < V_k$ be the selected levels of stimulus for the experiment; then $(V_1, V_2, \dots, V_k)$ is the selected design. Let $R(v)$ be the "true" response curve; the response curves considered here are cumulative distribution functions. Thus, the true probability of response p_i at stimulus V_i is $R(V_i)$. Next we generate a random variable, U_i , from a uniform distribution over $(0, 1)$ and set $X_i = 1$ if $U_i \leq p_i$, and $X_i = 0$ if $U_i > p_i$. Thus the outcome for a given design is a k -dimensional vector of 0's and 1's.

Once $\underline{X} = (X_1, X_2, \dots, X_k)$ is obtained, the probabilities of response, p_i 's, $i = 1, \dots, k$, can be estimated using the approach discussed in MS.

2.2 Estimation of $V_{.05}$

To estimate the probability of response, p_i , for each V_i , $i = 1, 2, \dots, k$, we assign a Dirichlet as a prior distribution for the successive differences $p_1, p_2 - p_1, \dots, p_k - p_{k-1}$ and the modal value of the joint posterior distribution is a Bayes point estimate of $(p_1, \dots, p_k)$. The computation of the modal value of the joint posterior distribution necessitates the use of an optimization algorithm; this is described by Mazzuchi and Soyer (1982).

The specification of the prior parameters of the Dirichlet distribution is also discussed in MS.

Once estimates of the p_i 's are obtained, an estimate of $V_{.05}$ can be obtained by constructing an estimated response curve. The estimated response curve is a plot of the levels of stimulus V_i , versus the $\hat{p}_i$'s, the estimated probabilities of response, $i = 1, 2, \dots, k$. Once such a plot is obtained, the interpolation procedure described in MS is used to estimate $V_{.05}$.

Specifically, for the estimation of $V_{.05}$, we first see if there is an observed stimulus, V_i , for which $\hat{p}_i = 0.05$. If so, then V_i is the estimate of $V_{.05}$. If not, the pair of observational stimuli, say V_i and V_{i+1} , for which $\hat{p}_i < 0.05 < \hat{p}_{i+1}$, are determined. Since the response curve is increasing, the straight line segment joining the points $0, \hat{p}_1, \dots, \hat{p}_i, \hat{p}_{i+1}, \dots, \hat{p}_k, 1$, will be an increasing function of i . We can find the value of the stimulus, say $\hat{V}_{.05}$, $V_i < \hat{V}_{.05} < V_{i+1}$, for which $\hat{p} = 0.05$ (as indicated in Figure 1).

2.3 Comparison of Designs

The goal of our analysis is to select a design, $(V_1, \dots, V_k)$, which will provide a "good" estimate of $V_{.05}$.

In order to determine an optimal choice of the k distinct levels of stimulus, we consider different designs, and first obtain an estimate of $V_{.05}$ for each design. Let $(V_1^j, \dots, V_k^j)$ denote design j ; the superscript j indicates a particular design. Once a j is chosen, we obtain $\underline{X}^j = (X_1^j, \dots, X_k^j)$ using the approach discussed in

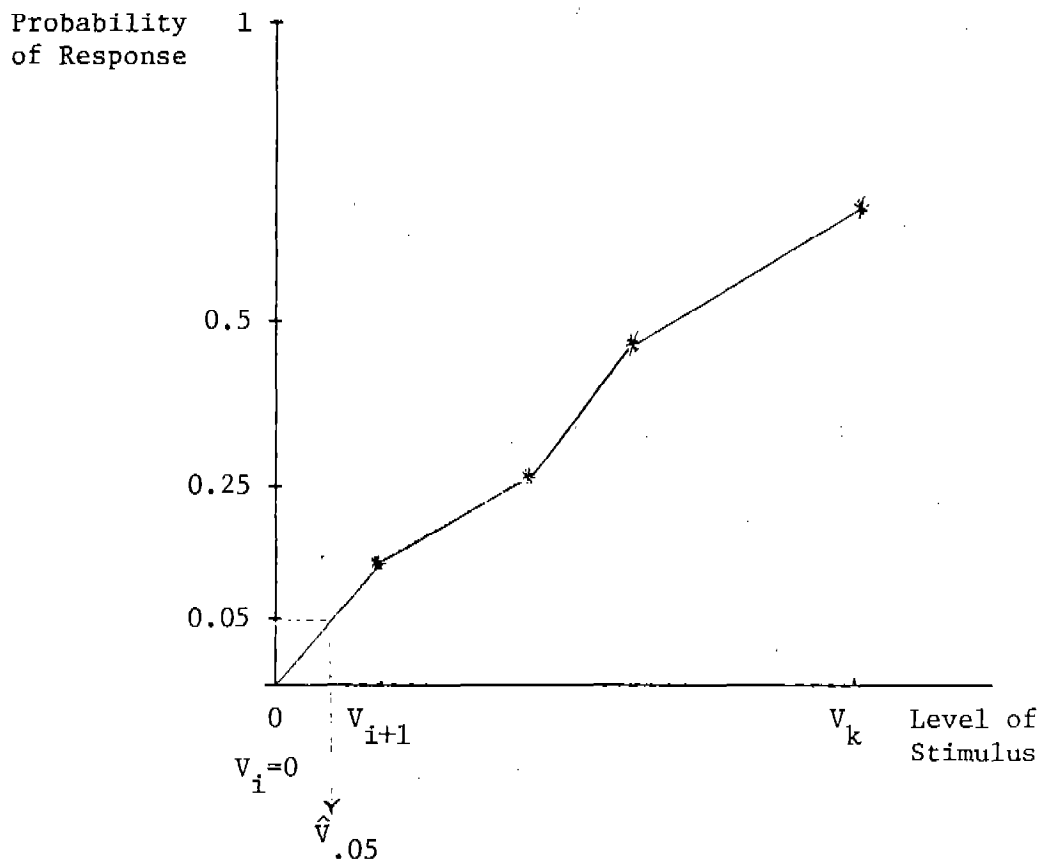


Figure 1. Interpolation procedure.

Section 2.2. Then via the estimated response curve, we obtain an estimate of $V_{.05}$, say $\hat{V}_{.05}^j$. Since the "true" response curve is assumed to be known, the estimate $\hat{V}_{.05}^j$ can be compared with the true value of $V_{.05}$.

If the above procedure is repeated for a different design, a different response curve is estimated. The various estimated response curves provide us with different estimates of $V_{.05}$, and we need to determine which of the designs gives us estimates which are closest to $V_{.05}$. Note that, since the outcome $\underline{X}^j = (X_1^j, \dots, X_k^j)$ for design j is obtained by simulation, different replications of $\underline{X}^j$ can be obtained by using different seeds in the simulation.

Let N be the number of replications which are analyzed for design j . For each replication of $\underline{x}^j$, a different response curve is estimated and therefore a different estimate of $V_{.05}$, say $\hat{V}_{.05}^j(\ell)$, is obtained. Since we know the true value of $V_{.05}$, the mean squared error (MSE) for design j is computed as

$$MSE^j = \sum_{\ell=1}^N (\hat{V}_{.05}^j(\ell) - V_{.05})^2.$$

The MSE for each design can be obtained and a comparison of the MSE's provides us with a criterion for selecting a good design. The design with the minimum MSE is a good design for a known response curve, say $R_i(v)$. It is possible that a design which is good for $R_i(v)$ may not be good for $R_k(v)$, $i \neq k$. This possibility has also been considered in our analysis.

3. SUMMARY

The approach we presented in Section 2 is applied to some simulated data in the next section.

Three different "true" response curves are selected. These curves are chosen in such a way that they will provide us with different values of $V_{.05}$.

The first response curve is specified via a Weibull distribution function,

$$R_1(v_1) = 1 - \exp\left\{-\left(\frac{v_1}{100}\right)^2\right\}, \text{ where } V_{.05} = 22.$$

The second is via a lognormal distribution function,

$$R_2(v_i) = \Phi \left(\frac{\log_e v_i - 4.50}{0.33} \right), \text{ where } V_{.05} = 52$$

and Φ denotes the standard normal distribution function.

The third response curve considered is also a lognormal distribution function, which gives $V_{.05} = 10$; that is,

$$R_3(v_i) = \Phi \left(\frac{\log_e v_i - 3.3}{0.6} \right).$$

Five different designs are selected and analyzed.

Design 1 -- the k observations are distributed evenly over the entire interval of the range of testing, say I .

Design 2 -- all the k observations are concentrated on the left-hand half of I .

Design 3 -- all the k observations are concentrated in the center of I .

Design 4 -- all the k observations are concentrated on the right-hand half of I .

Design 5 -- the k observations are sequentially obtained in three different phases.

The value of k is (arbitrarily) chosen as 12, and due to the expense of simulation, ten different replications of $\underline{X}^j$ are considered.

The MSE's for each design based on the ten replications are computed, on the basis of which it is felt that Design 3 is a suitable design for the estimation of $V_{.05}$.

4. APPLICATION TO SOME SIMULATED DATA

The three "true" response curves discussed in Section 3 are analyzed separately in this section. These response curves are illustrated in Figures 2, 3, and 4. A replication, simulated from each of these response curves, is presented in the Appendix for illustration.

We assume that the probability of a response at a striking velocity of 300 is almost 1. Thus we make an arbitrary choice for our best prior guess of p_1 , say p_1^* , by letting $p_1^* = 1 - \exp[-0.0307 V_1]$. The prior parameters are chosen as described in MS. In our analysis the smoothing parameter is chosen as $\beta = 10$.

The five different designs presented in Section 3 will be used in the analysis. In the first four designs, the penetrator is tested in a single phase. In Design 1, the 12 observations are taken equally spaced over the entire range of testing, (0,300). In Design 2 all 12 observations are taken equally spaced on the left-hand half of the interval (0,150). In Designs 3 and 4 the 12 observations are taken equally spaced in the center, and on the right-hand half of the interval, respectively.

The sequential design, Design 5, consists of three phases. In the first phase, six observations are taken equally spaced over the entire range of testing. Ten different replications of the outcome vector, $\underline{X}$, are examined and the experimenter tries to identify two regions: one region where the outcome is zero and another where the outcome is one most of the time. Once these two regions are determined, the experimenter has knowledge about the region where the response is

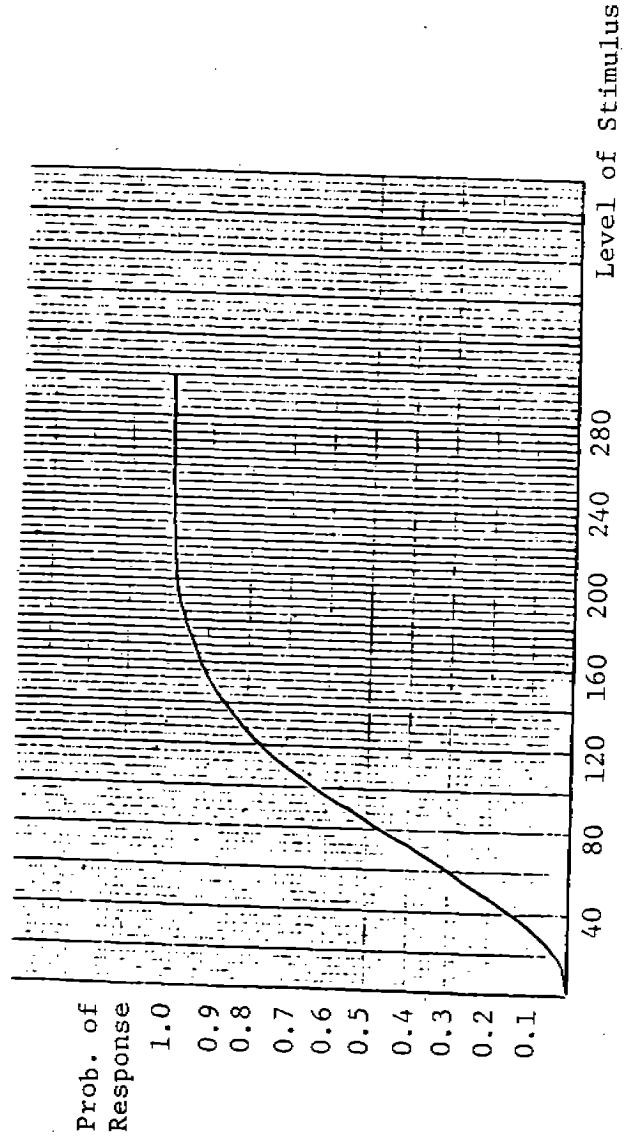


Figure 2. Weibull response curve.

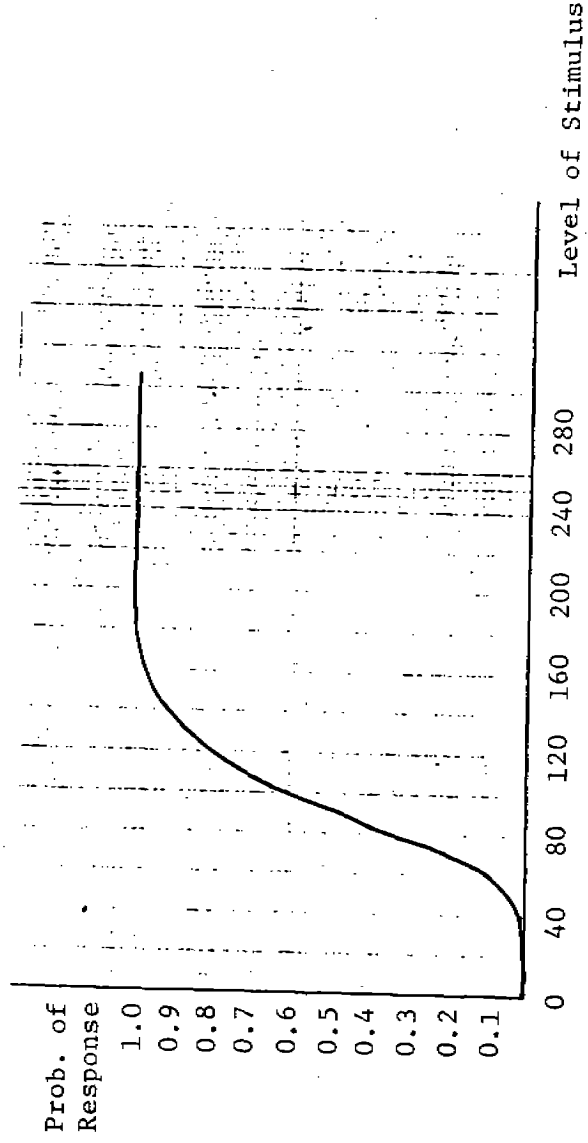


Figure 3. Lognormal response curve I.

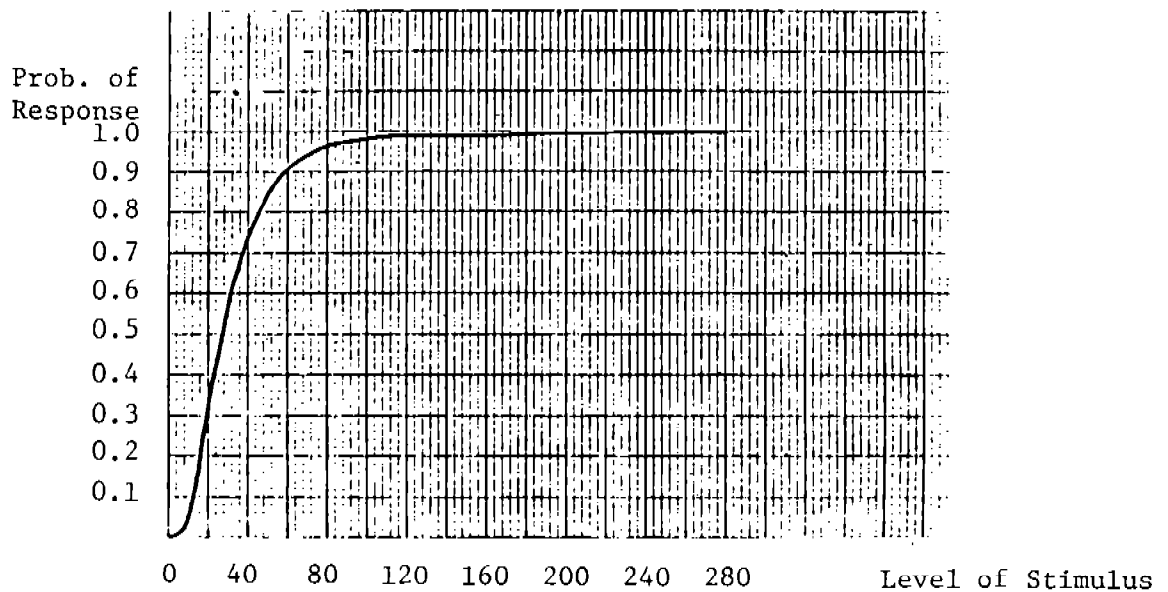


Figure 4. Lognormal response curve II.

most likely to change. In phase 1 of the experiment a new response curve is also estimated, and this curve provides the prior values of p_i 's for the second phase. In the second phase, three observations are taken, equally spaced over the region where the response is most likely to change, as is indicated by phase 1 and a new response curve estimated based on these observations and the prior (the posterior from phase 1). In the final phase of the experiment, the remaining three observations are taken on the left-hand end of our best guess based on phase 2 and a new response curve is estimated using these observations and the prior (the posterior from phase 2). The estimate of $V_{.05}$ is obtained by using this updated response curve.

The outcomes of the five different designs are presented in the Appendix, Tables A.1 - A.3.

4.1 Analysis for the Weibull Response Curve

The first response curve that is considered is a Weibull distribution function for which $V_{.05} = 22$. The outcome vector, $\underline{x}^j$ for $j = 1, \dots, 4$, is simulated using this response curve. Ten replications of the outcome vector are obtained for each design. One of these replications is presented in Table A.1 of the Appendix. The procedure that was discussed in Section 2.2 is adopted and the estimates of $V_{.05}$ are obtained. The "true" response curve and the estimated response curve are plotted in Figures 5, 6, 7, and 8 for one replication, and presented in Table A.1. The estimates of $V_{.05}$ are obtained from these figures. For Design 1, the estimate of $V_{.05}$ is obtained as $\hat{V}_{.05}^1 = 4$ from the estimated response curve in Figure 5. Similarly, the estimates of $V_{.05}$ for Designs 2, 3, and 4 are obtained as $\hat{V}_{.05}^2 = 4$, $\hat{V}_{.05}^3 = 10$, and $\hat{V}_{.05}^4 = 16$.

For the sequential design, the response curve that is estimated in the first phase is presented in Figure 9 for the replication presented in Table A.1. The response curves estimated in phases 2 and 3 are plotted in Figures 10 and 11, respectively. The estimate of $V_{.05}$ is obtained from Figure 11 as $\hat{V}_{.05}^5 = 8$.

Once the $\hat{V}_{.05}^j(\ell)$ values are obtained for $\ell = 1, \dots, 10$ for Design j , MSE^j can be computed as:

$$MSE^j = \frac{1}{10} \sum_{\ell=1}^{10} \left(\hat{V}_{.05}^j(\ell) - 22 \right)^2 .$$

MSE^j 's for the Weibull response curve are presented in Table 1.

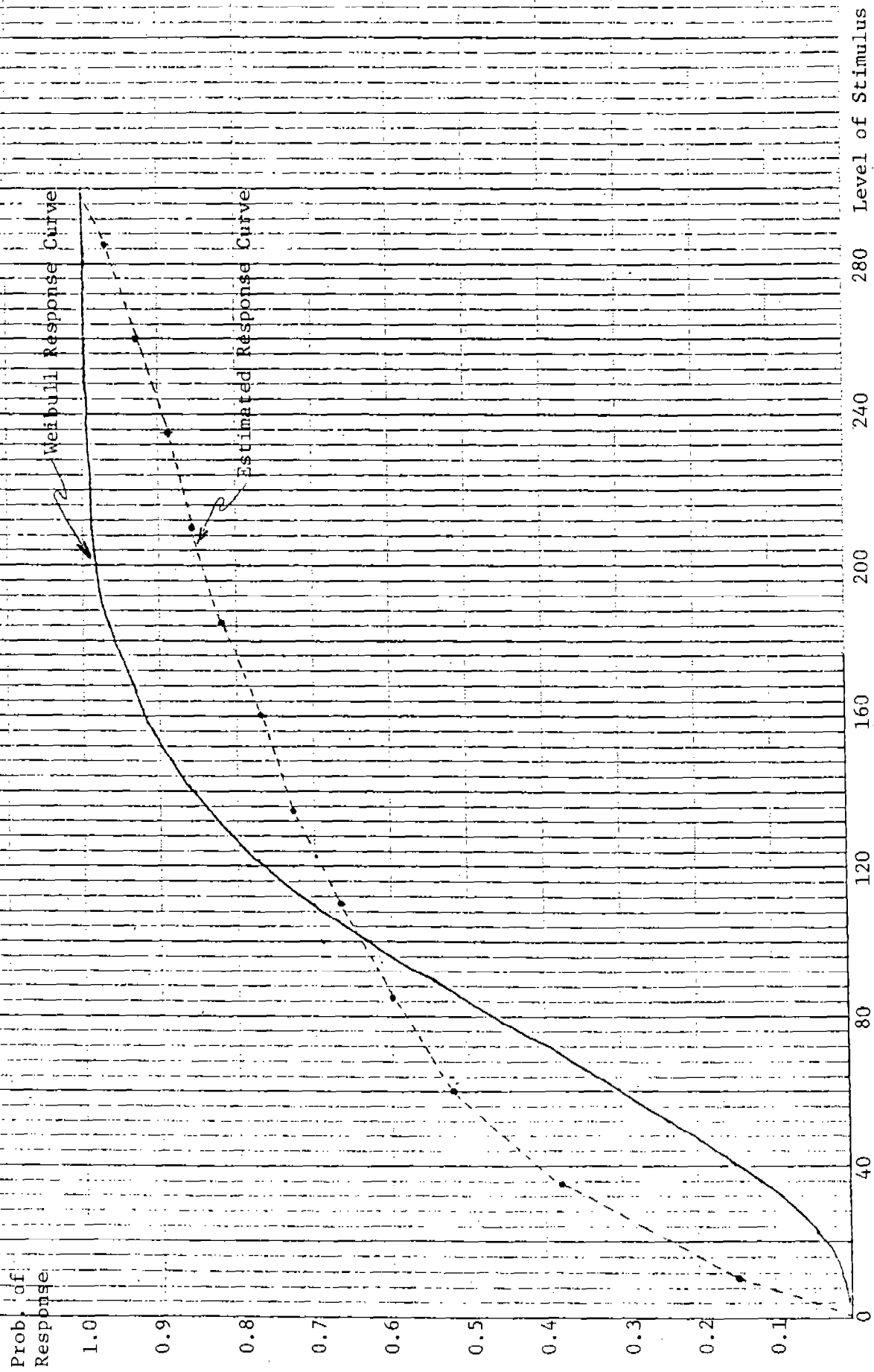


Figure 5. Design 1.

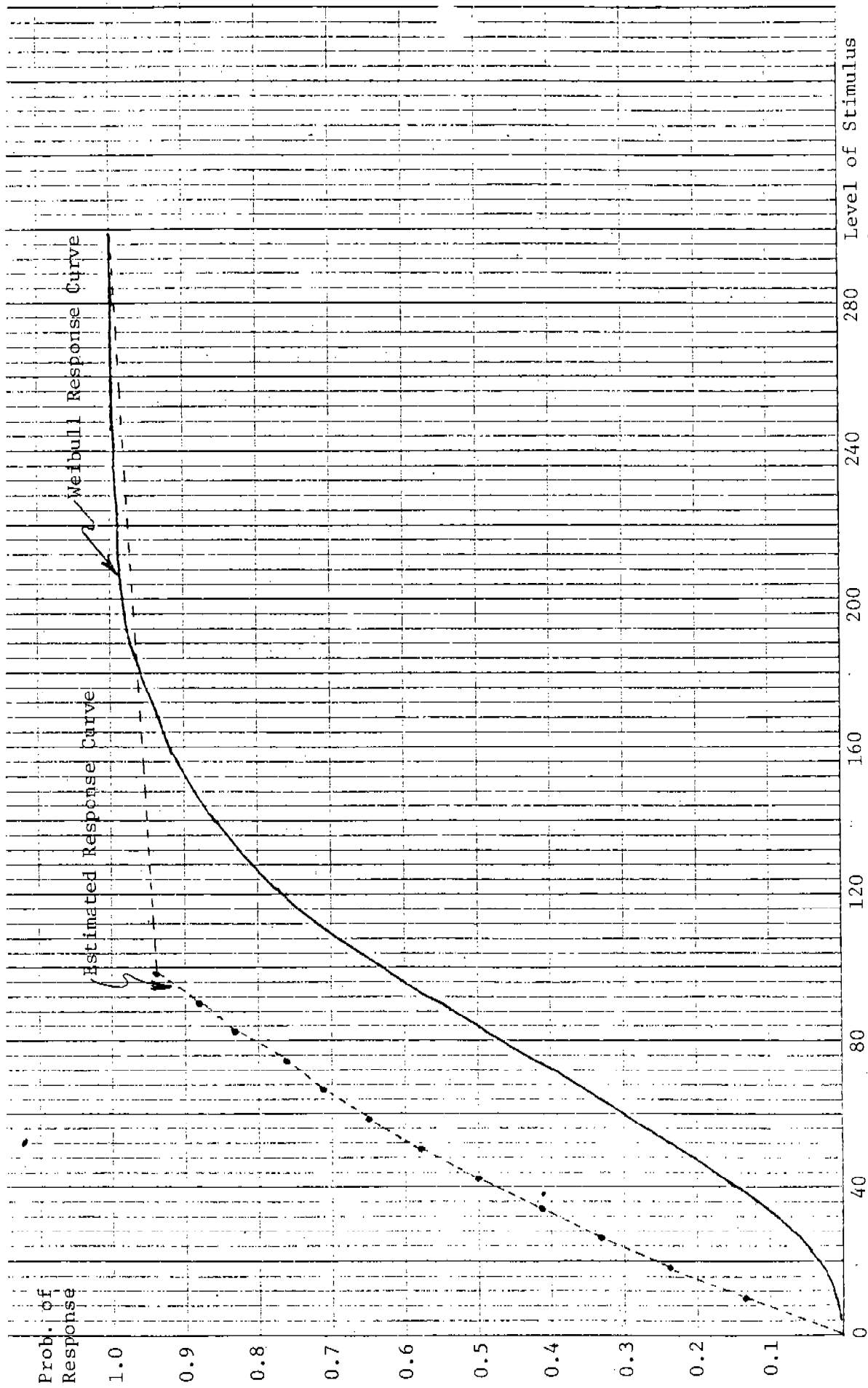


Figure 6. Design 2.

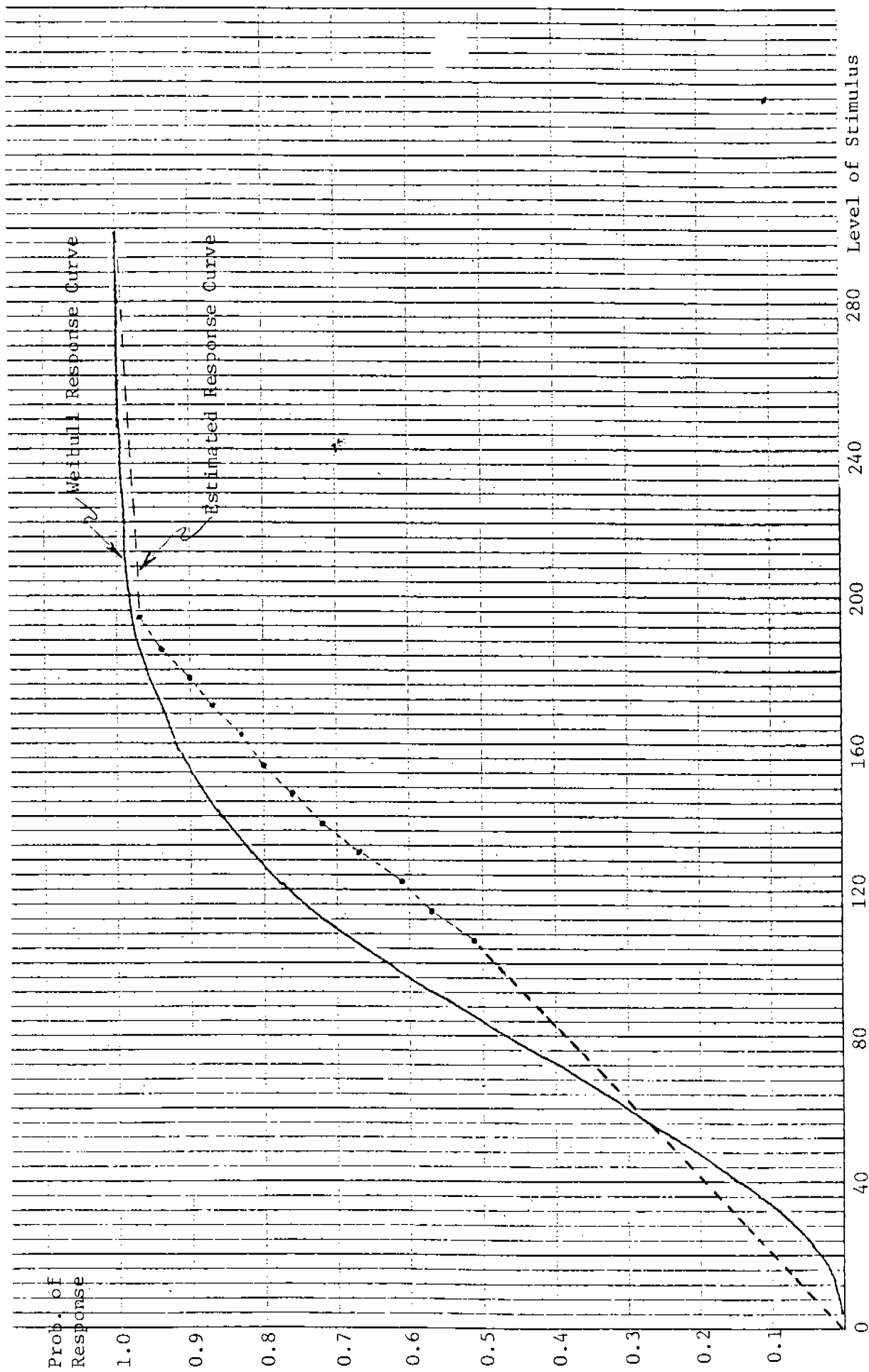


Figure 7. Design 3.

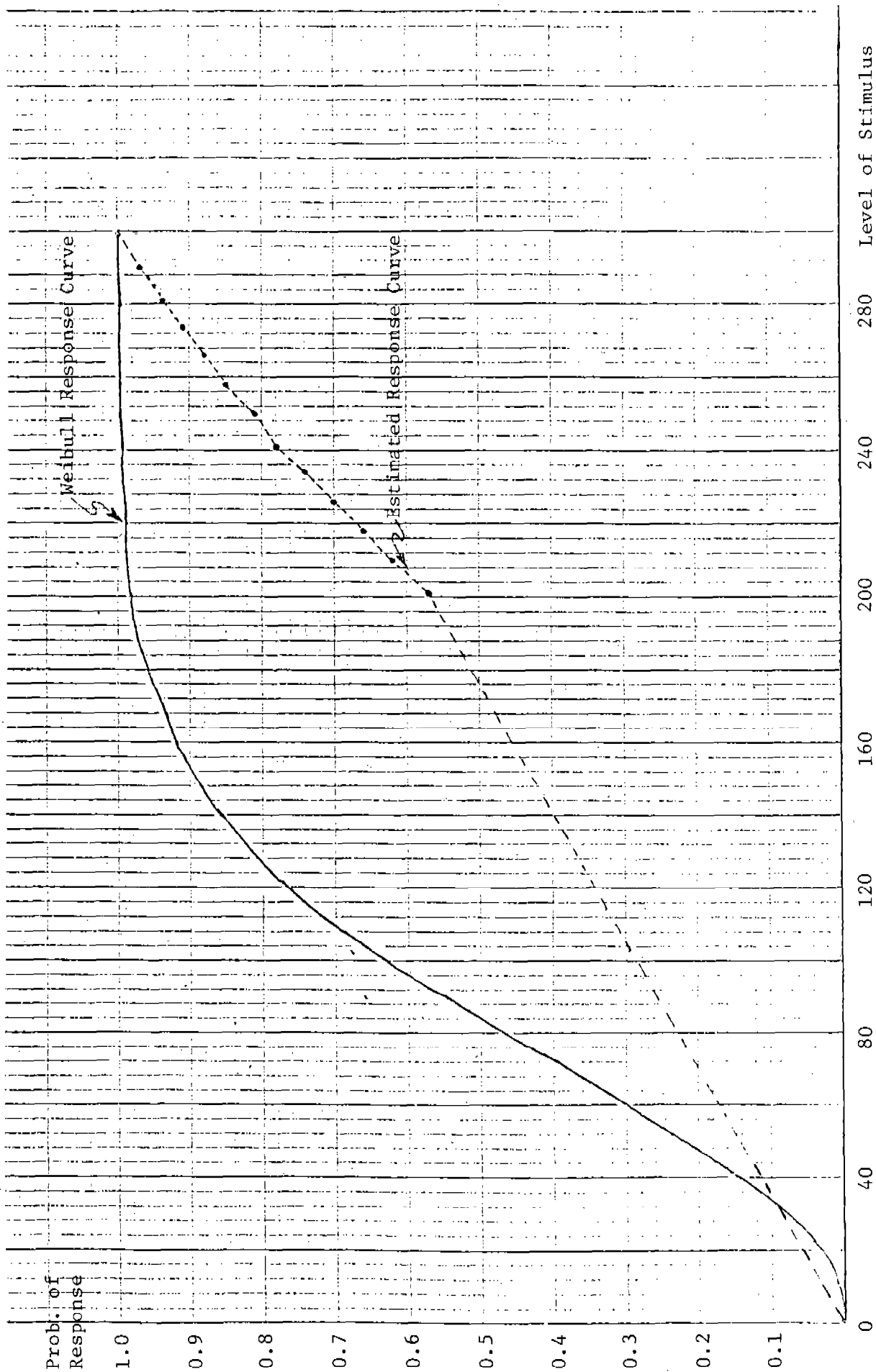


Figure 8. Design 4.

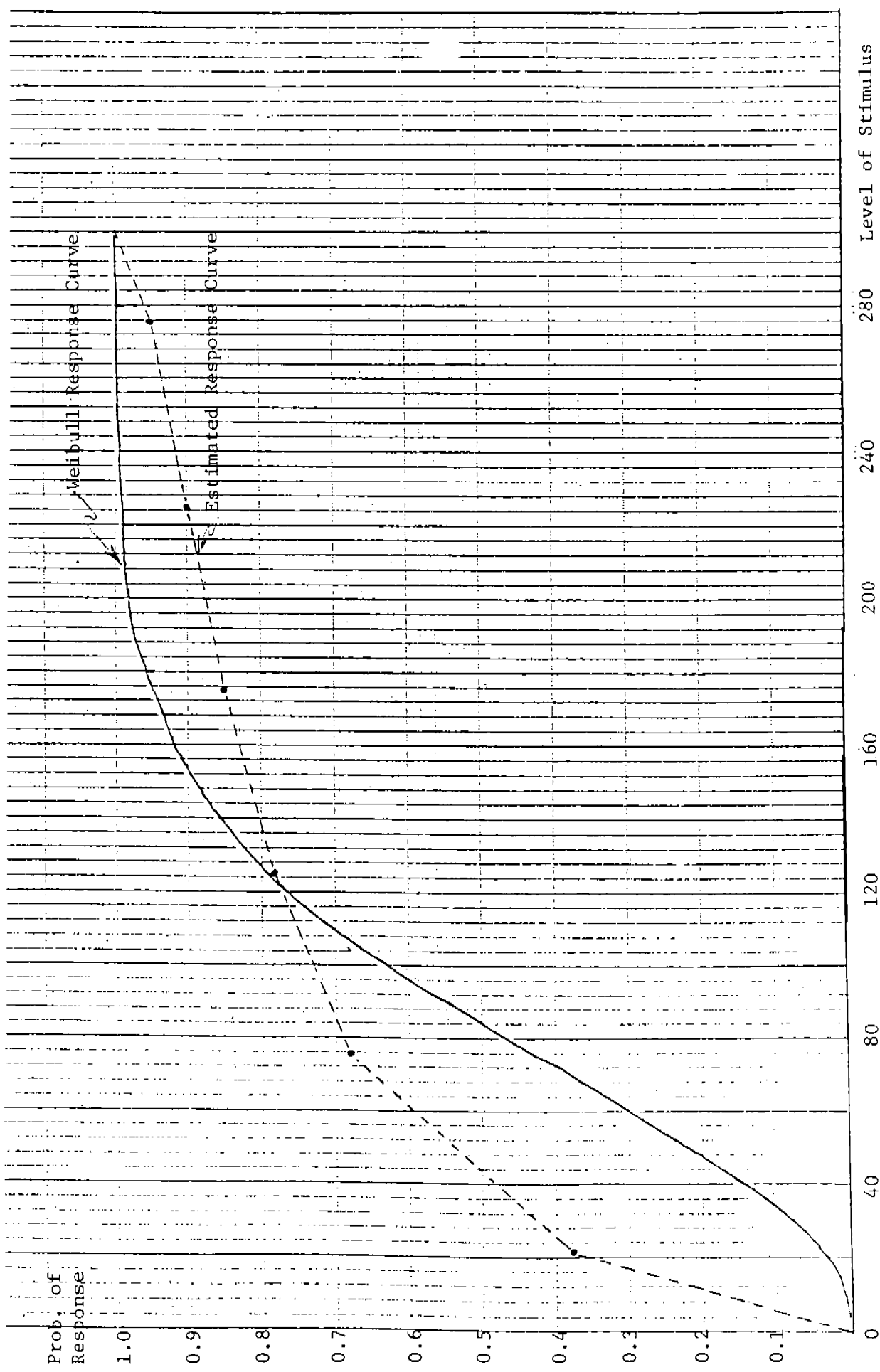


Figure 9. Sequential design, phase I.

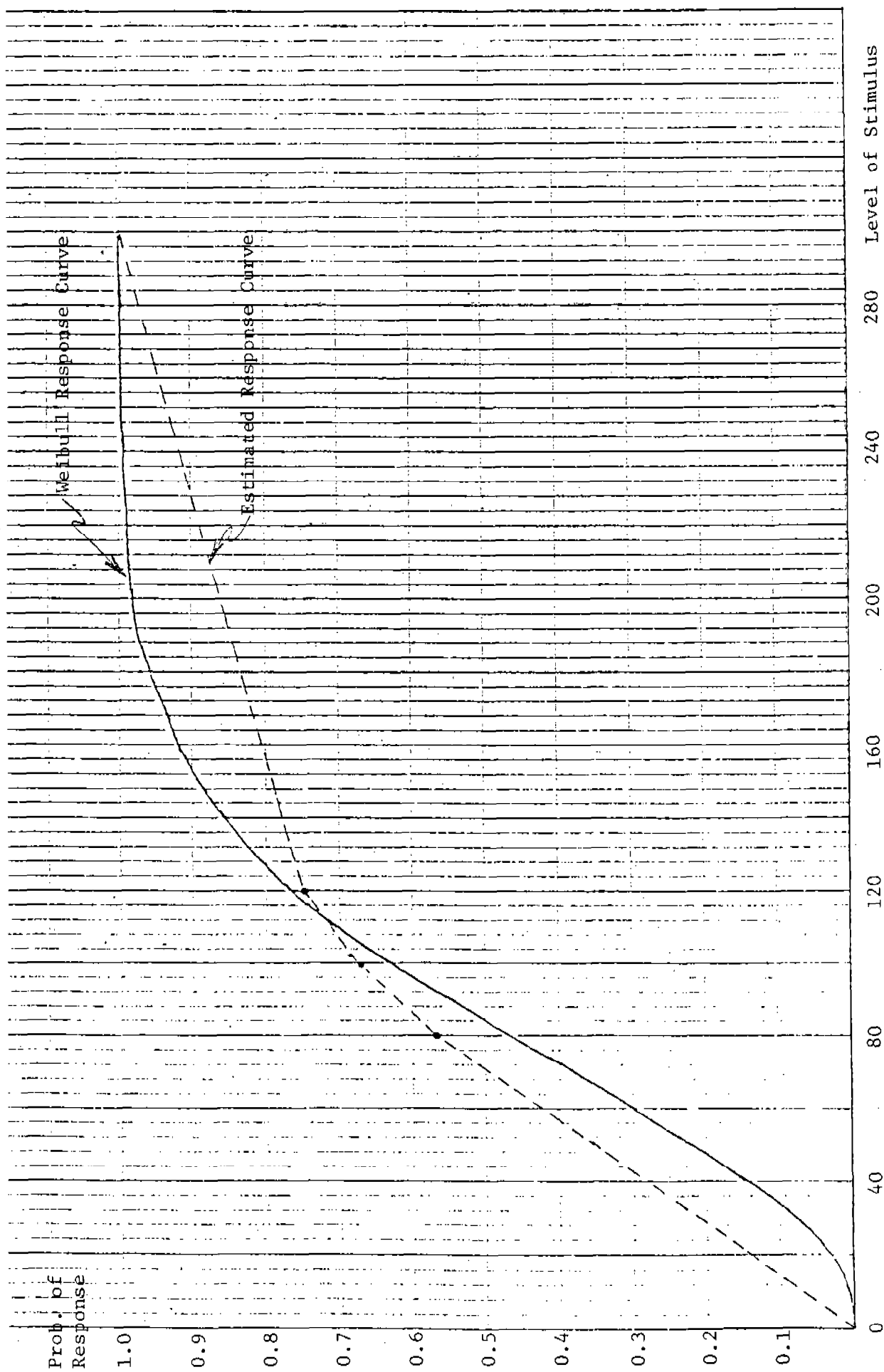


Figure 10. Sequential design, phase II.

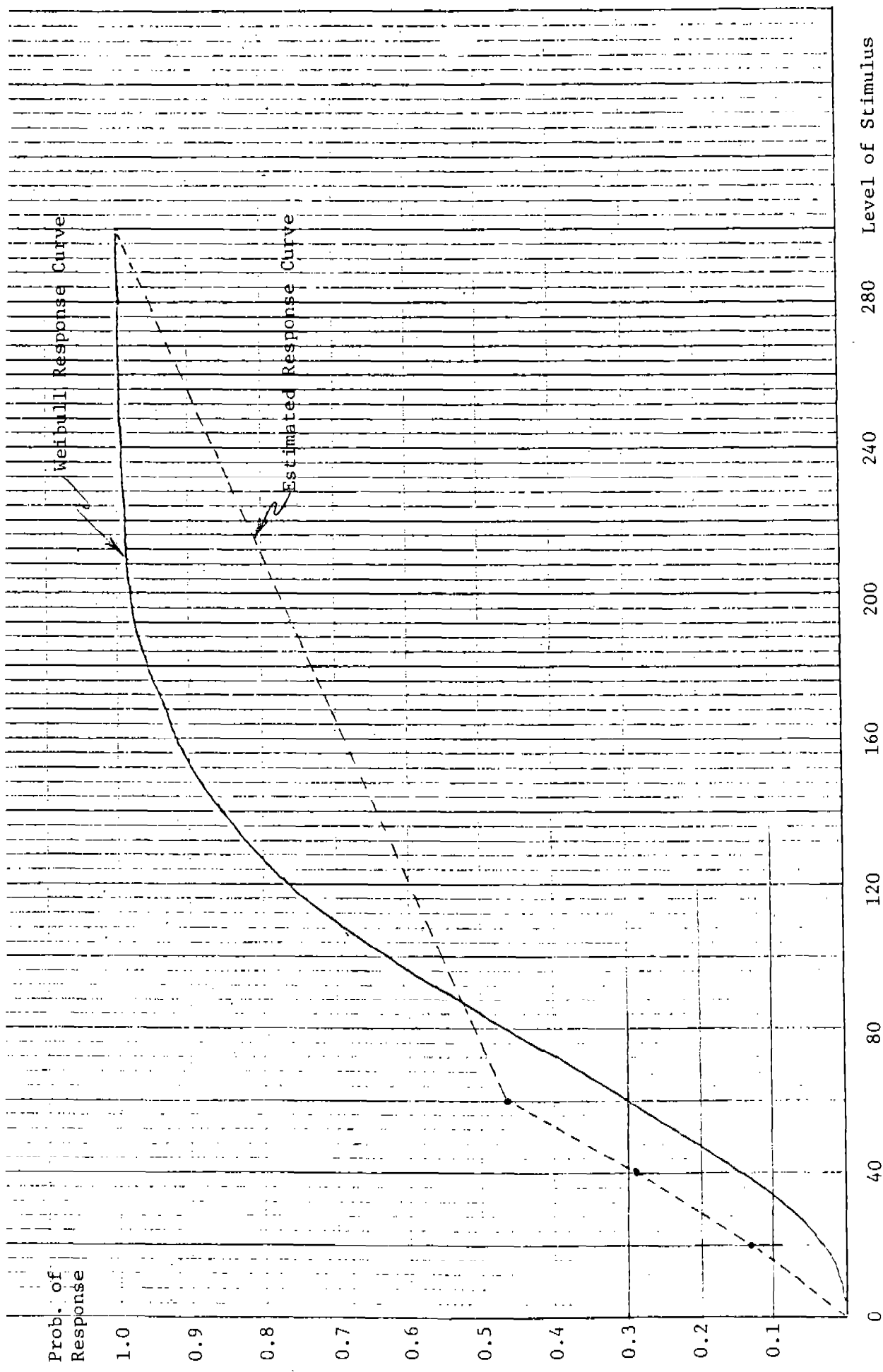


Figure 11. Sequential design, phase III.

Table 1. MSE's for the Weibull Response Curve

Design	MSE
1	360.1
2	320.2
3	126.2
4	36.0
5 (sequential)	196.8

The minimum MSE is obtained when Design 4 is selected; that is, when all k observations are concentrated in the right-hand half of the interval of the range of testing. The second lowest MSE is obtained when all k observations are concentrated in the center of the interval of the range of testing.

4.2 Analysis for the Lognormal Response Curve I

The second response curve is a lognormal distribution function where $V_{.05} = 52$. Again ten outcome vectors, $\underline{X}^j$'s, are simulated for each design. The response curves are constructed and $\hat{V}^j$'s are obtained. The estimated response curves can be observed from Figures 12, 13, 14, and 15 for Designs 1, 2, 3, and 4, respectively, for a single replication. The estimates of $V_{.05}$ are $\hat{V}_{.05}^1 = 4$, $\hat{V}_{.05}^2 = 5$, $\hat{V}_{.05}^3 = 11$, and $\hat{V}_{.05}^4 = 18$ from the corresponding figures.

For the sequential design, Design 5, the response curves estimated in phases 1, 2, and 3 are plotted in Figures 16, 17, and 18, respectively. The estimate of $V_{.05}$ is obtained as $\hat{V}_{.05}^5 = 9$ from Figure 18.

Prob. of
Response

1.0

0.9

0.8

0.7

0.6

0.5

0.4

0.3

0.2

0.1

Lognormal Response Curve I

Estimated Response Curve

0

40

80

120

160

200

240

280

Level of Stimulus

Figure 12. Design 1, lognormal.

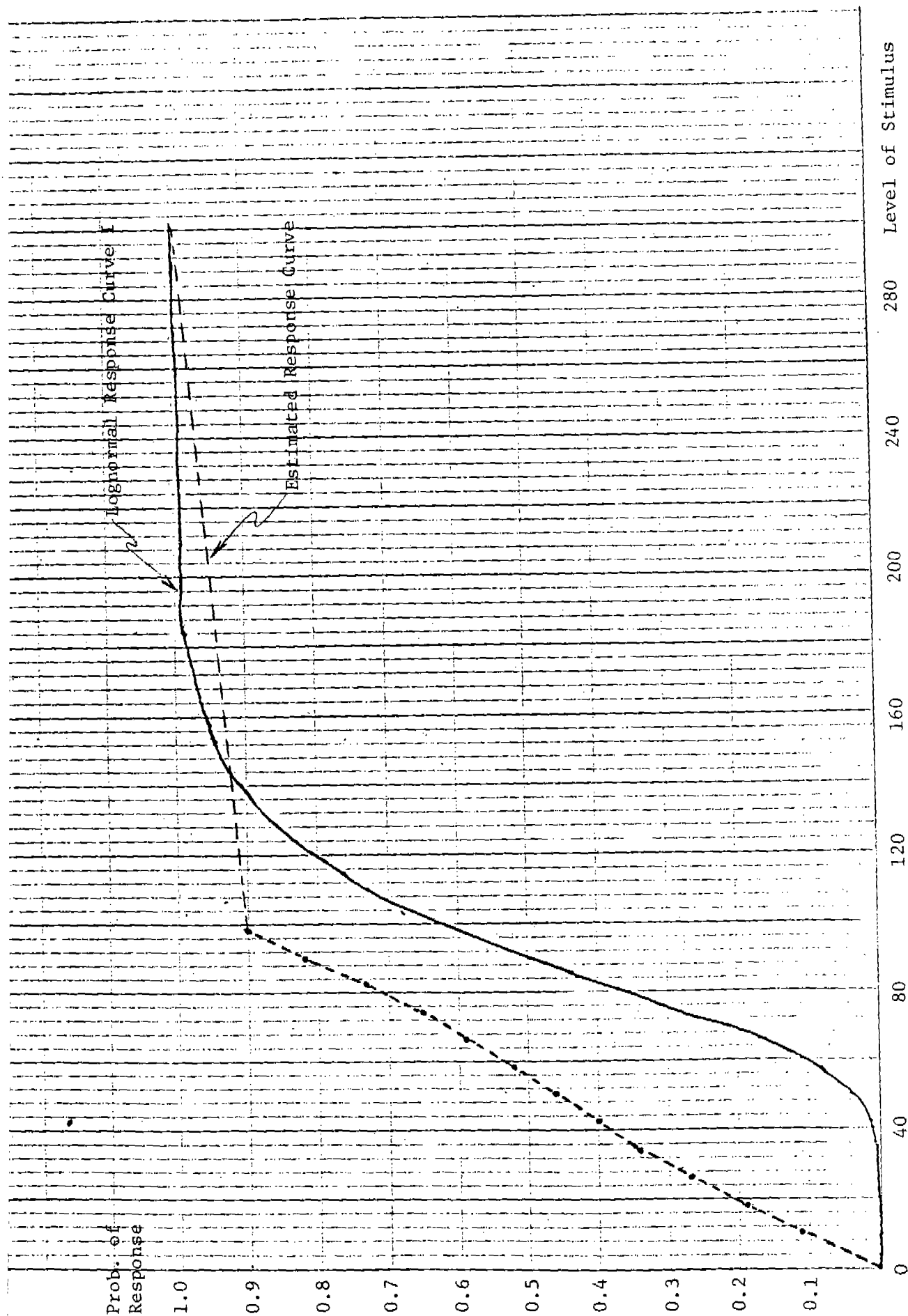


Figure 13. Design 2, lognormal.

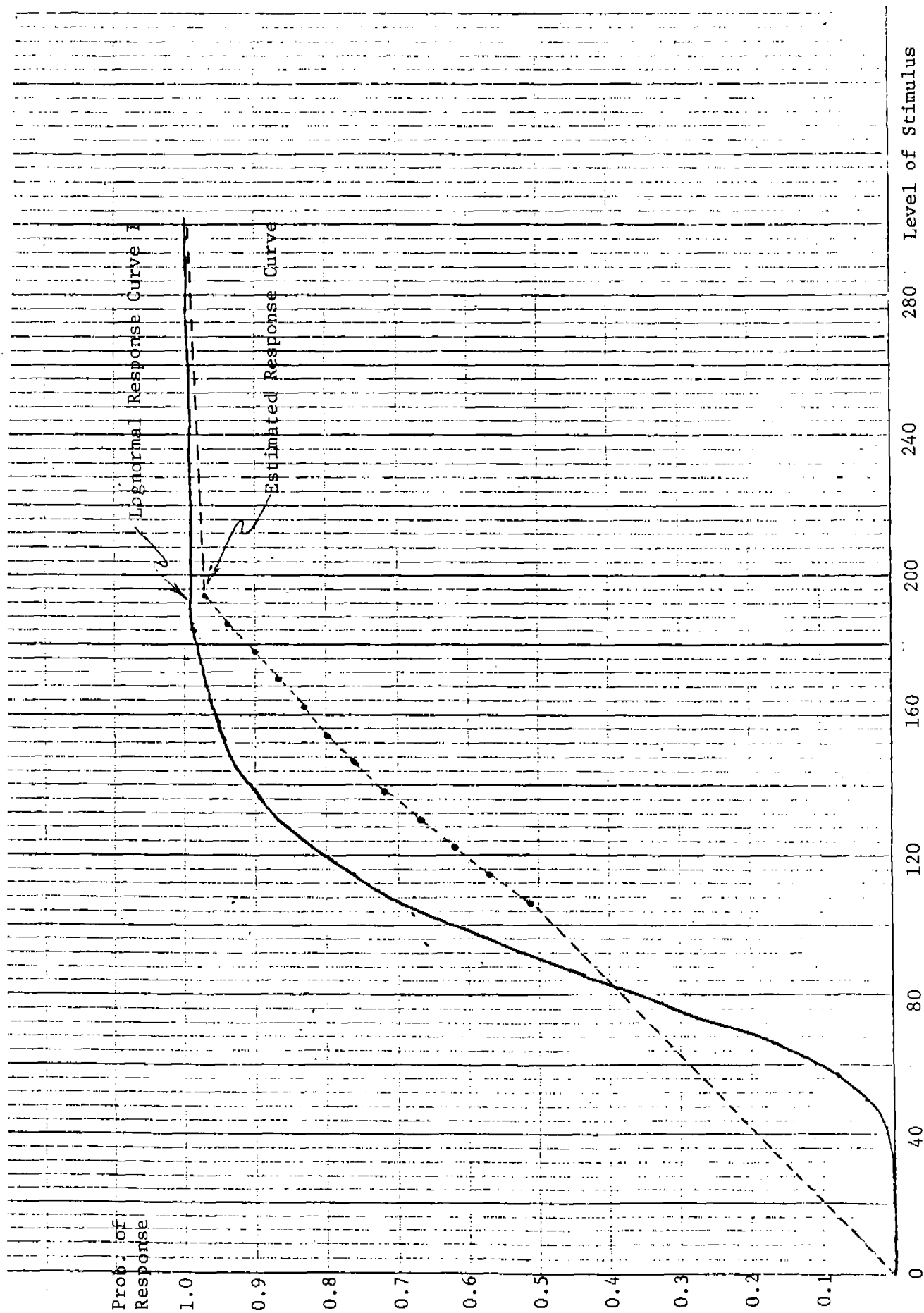


Figure 14. Design 3, lognormal.

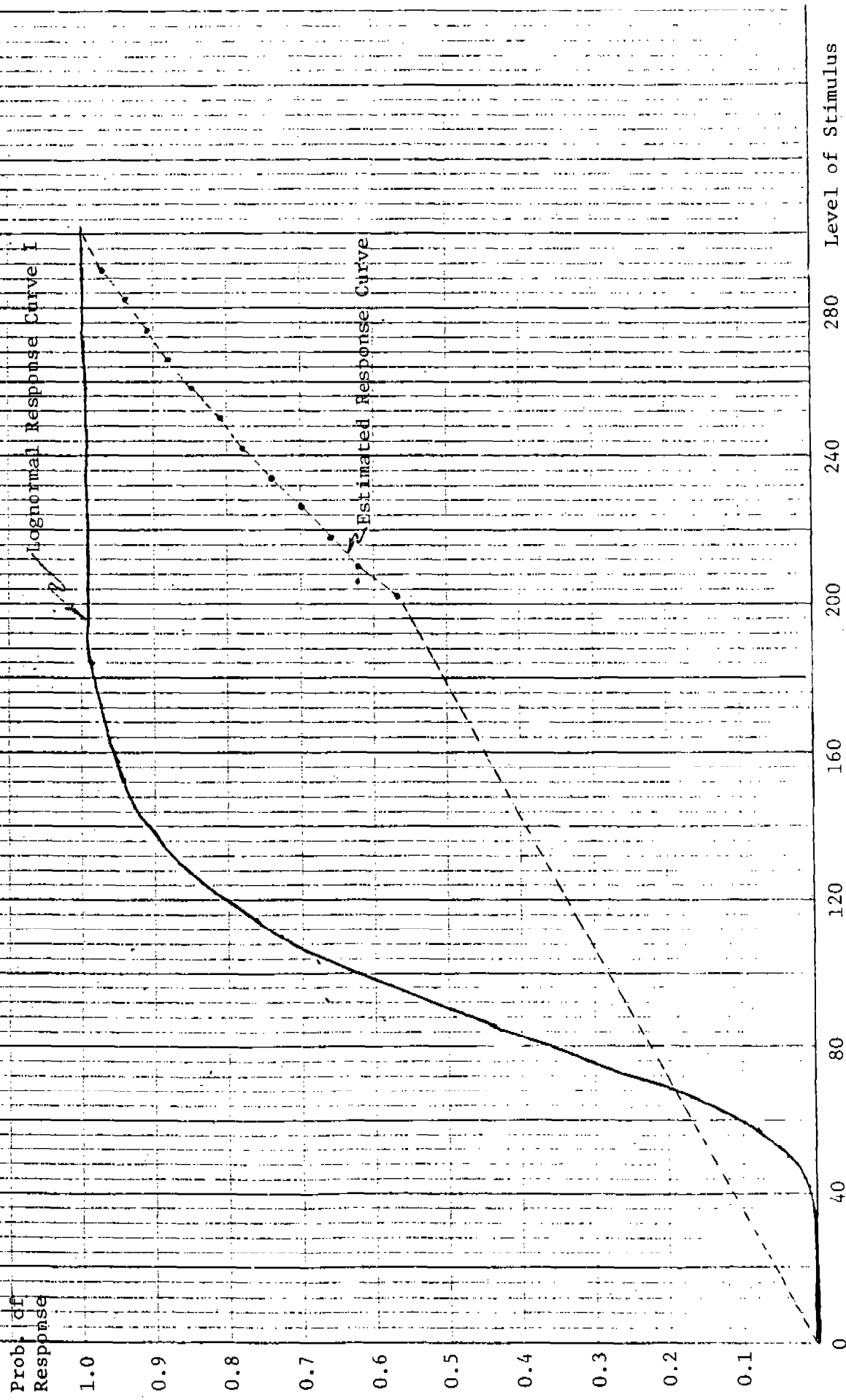


Figure 15. Design 4, lognormal.

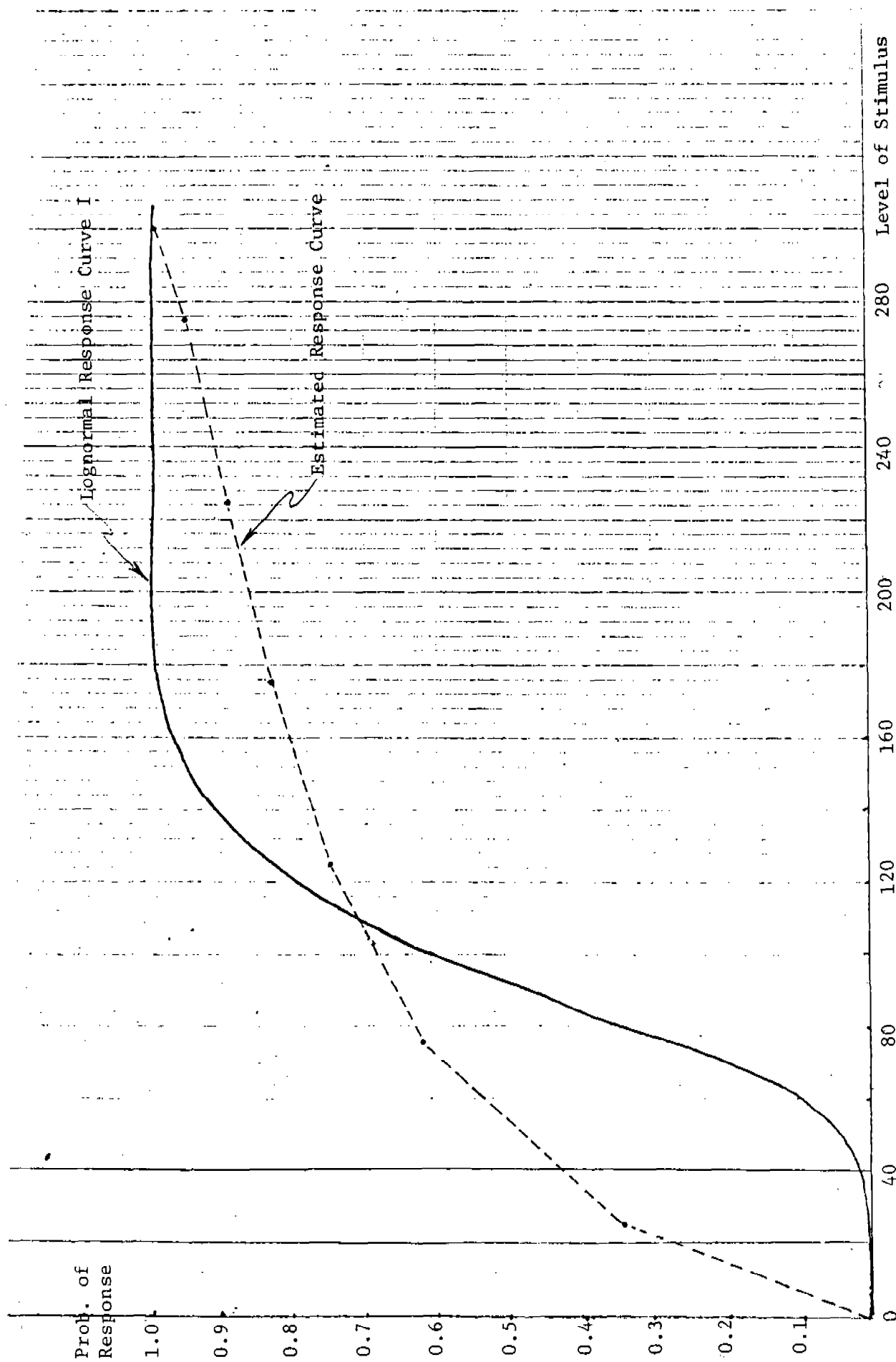


Figure 16. Sequential design, phase I, lognormal.

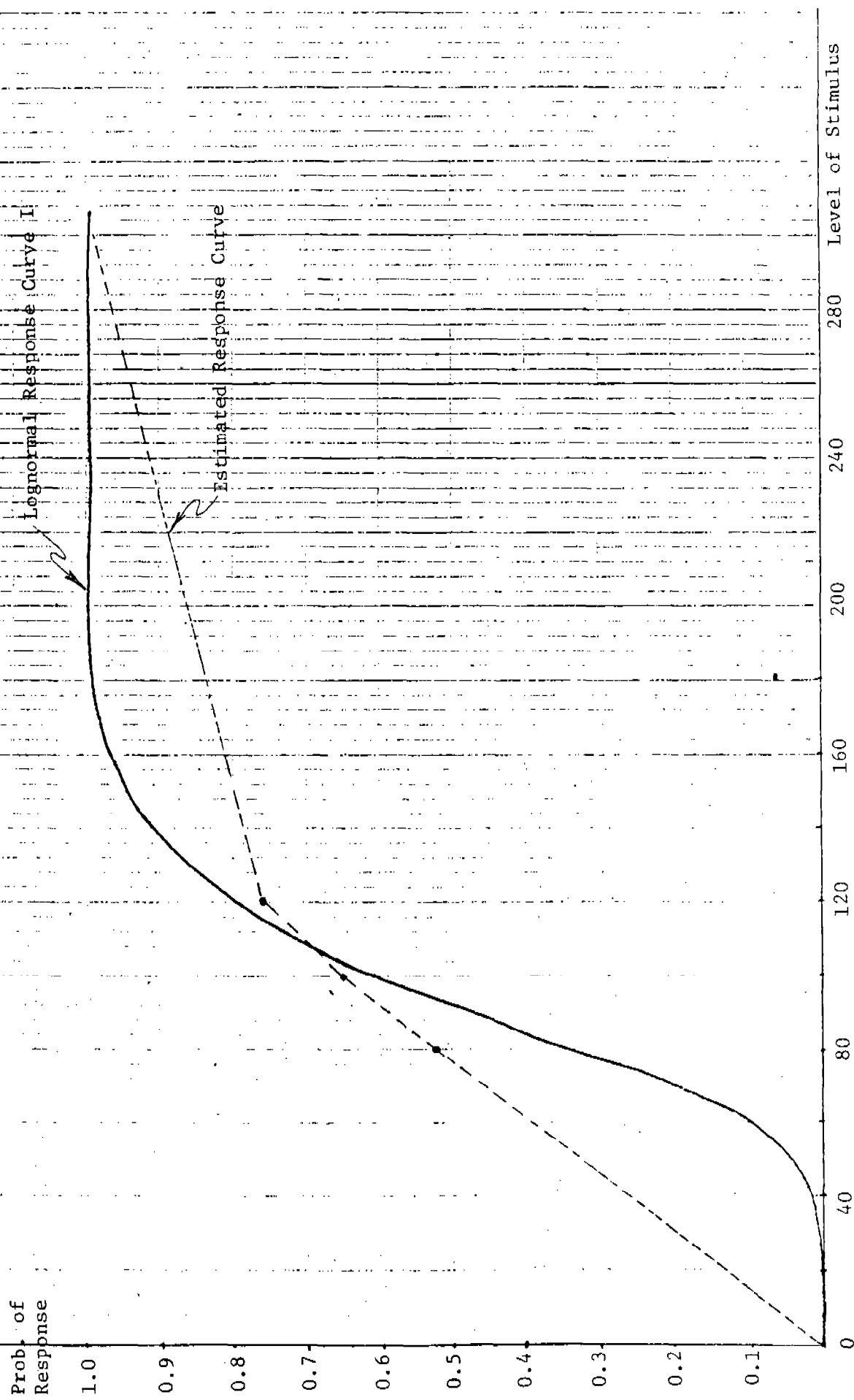


Figure 17. Sequential design, phase II, lognormal.

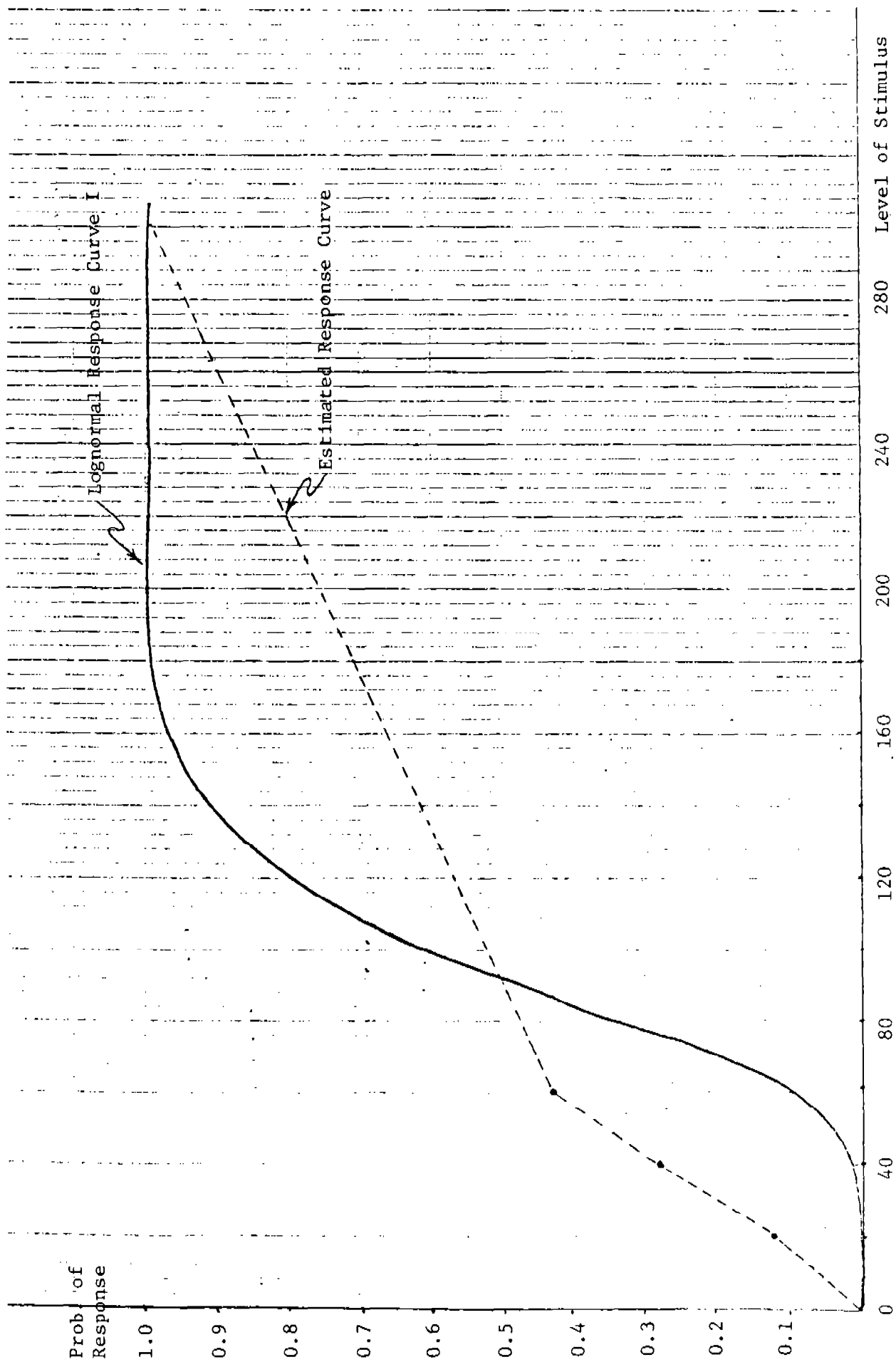


Figure 18. Sequential design, phase III, lognormal.

Table 2. MSE's for the Lognormal Response Curve I

Design	MSE
1	2313.7
2	2209.0
3	1714.2
4	1142.8
5	1981.1

The MSE's computed for the lognormal response curve are presented in Table 2. As we can observe, the minimum MSE is obtained when Design 4 is selected. The second lowest MSE is obtained for Design 3.

4.3 Analysis for the Lognormal Response Curve II

The third response curve is also a lognormal distribution function, where $V_{.05} = 10$. The outcome vectors are simulated and the response curves are estimated as in the previous sections. The $\hat{V}_{.05}^j$ values are obtained using the estimated response curves for each design. The estimated response curves can be observed from Figures 19 - 22 for Designs 1, 2, 3, and 4 for a single replication. The estimates are obtained as $\hat{V}_{.05}^1 = 2$, $\hat{V}_{.05}^2 = 3$, $\hat{V}_{.05}^3 = 10$, and $\hat{V}_{.05}^4 = 19$.

For Design 5, the estimated response curves for phases 1 - 3 are presented in Figures 23 - 25. The estimate of $V_{.05}$ is obtained as $\hat{V}_{.05}^5 = 3$ for the sequential design.

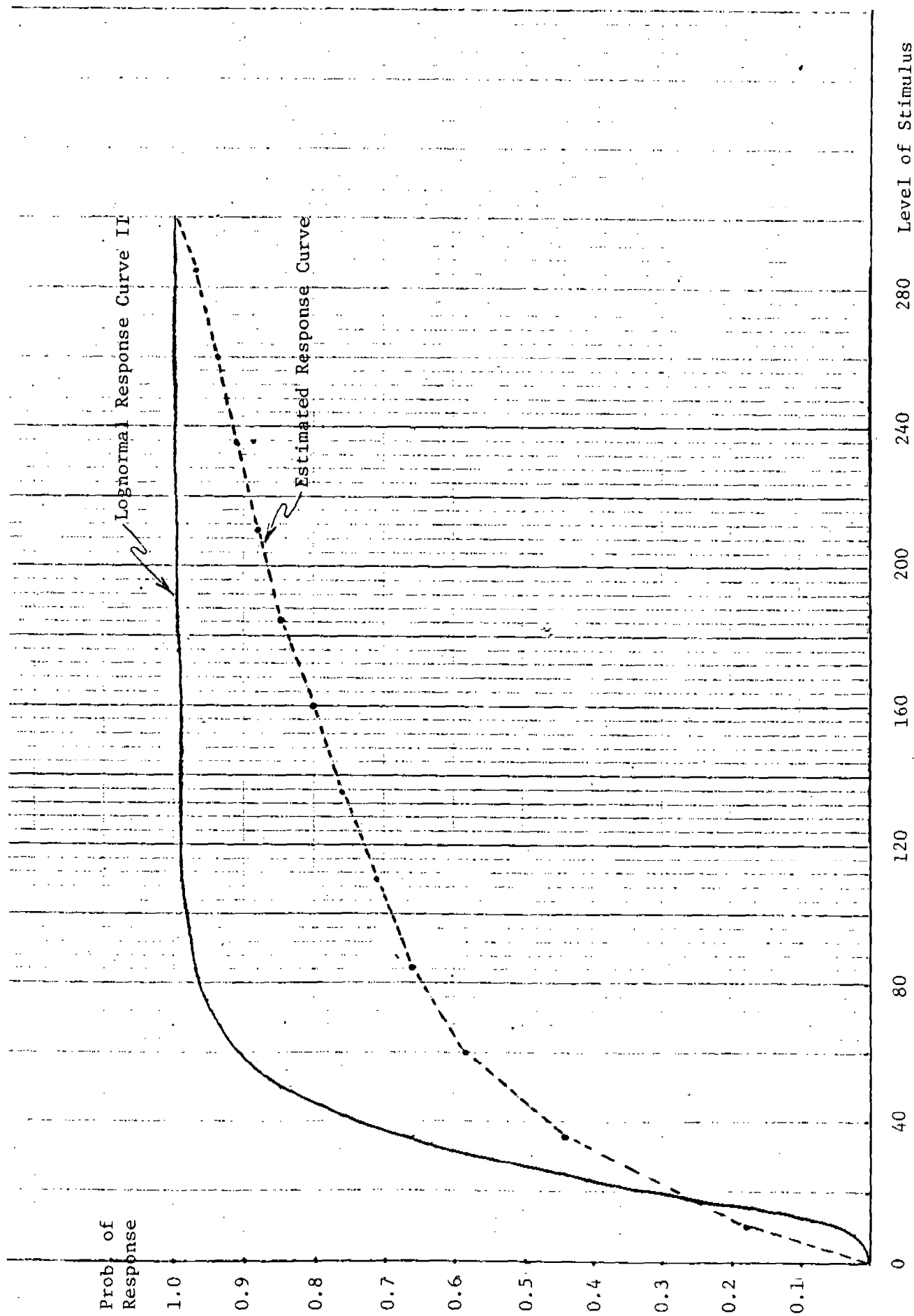


Figure 19. Design 1. lognormal II

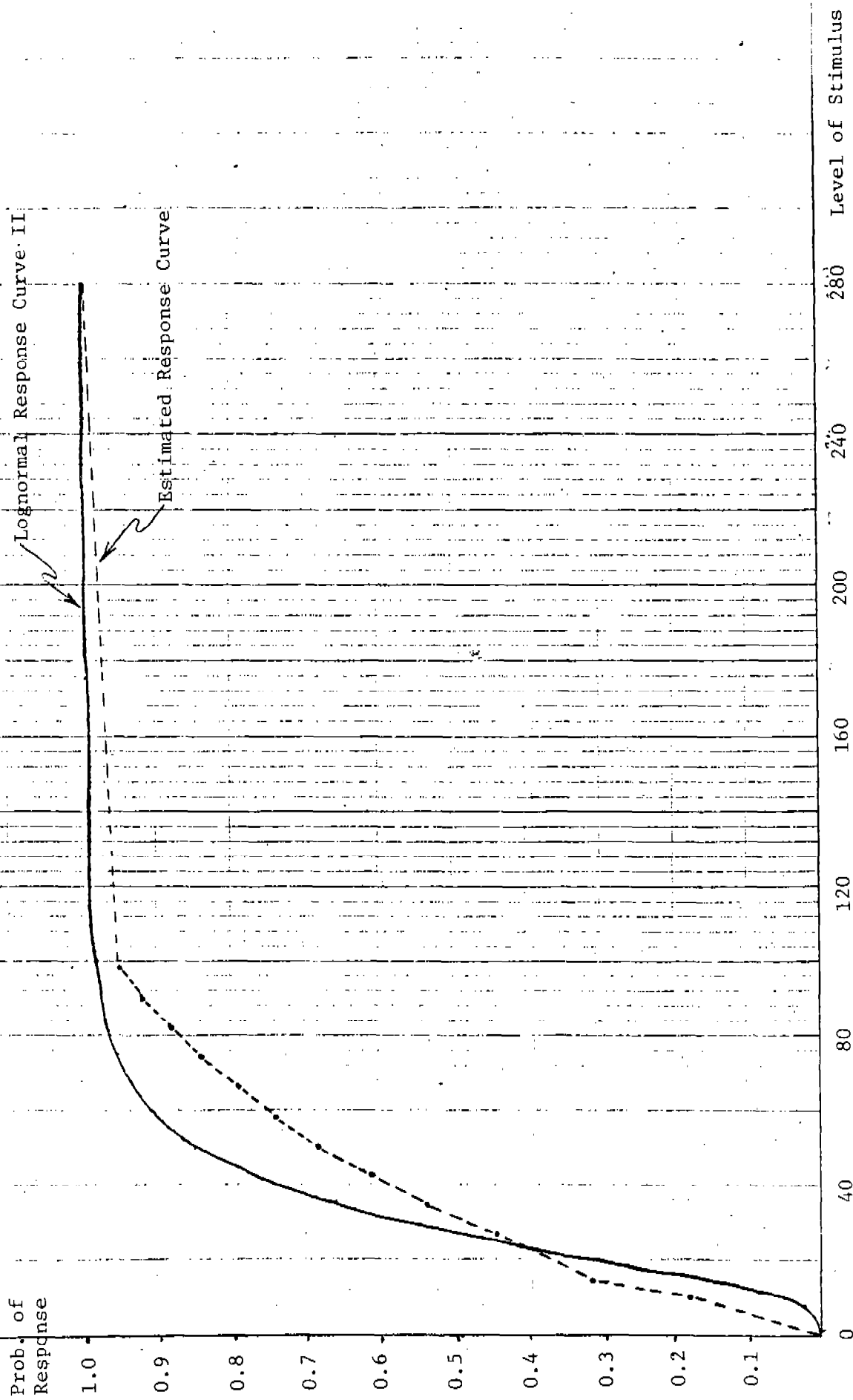


Figure 20. Design 2, lognormal II.

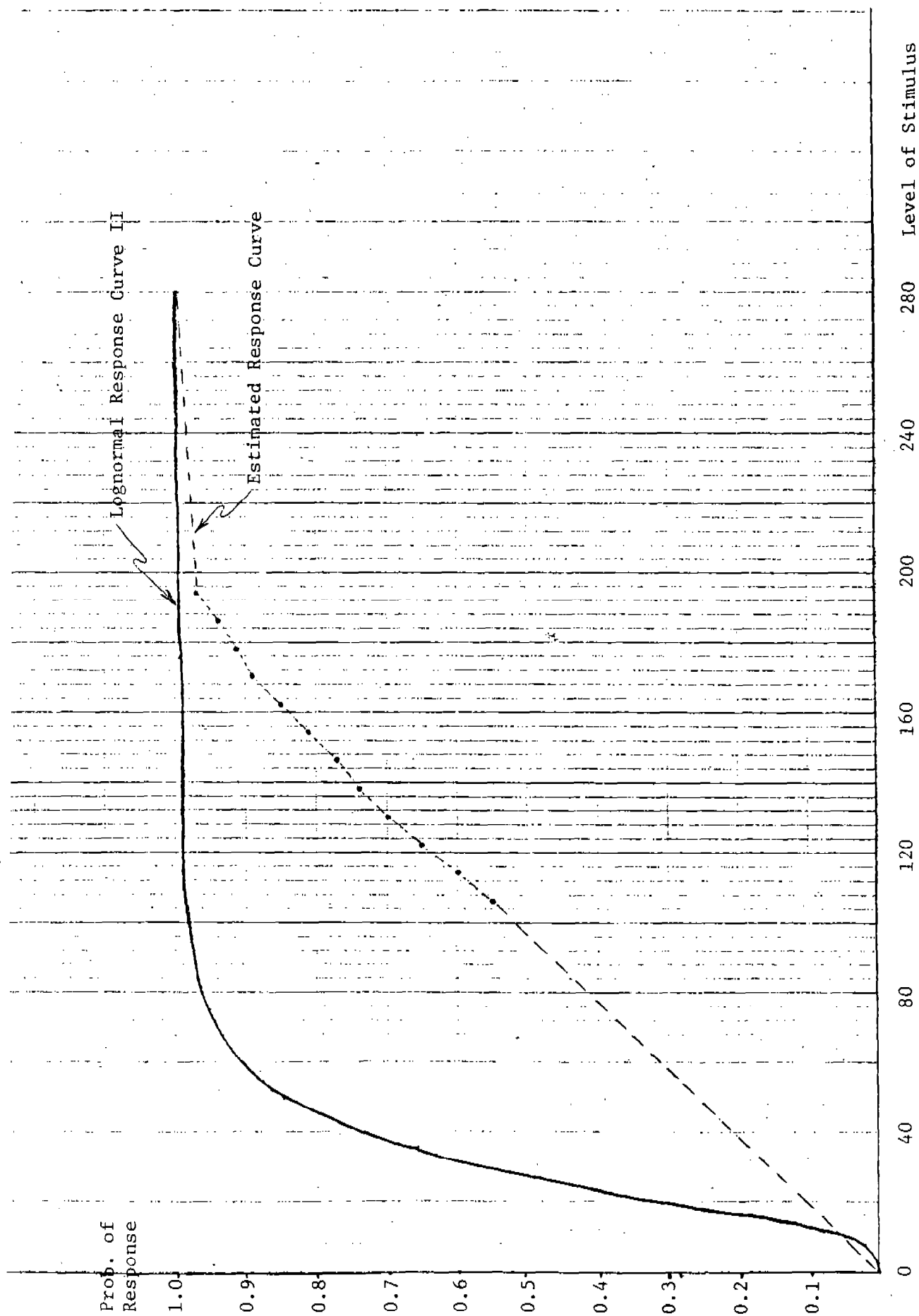


Figure 21. Design 3. Lognormal II

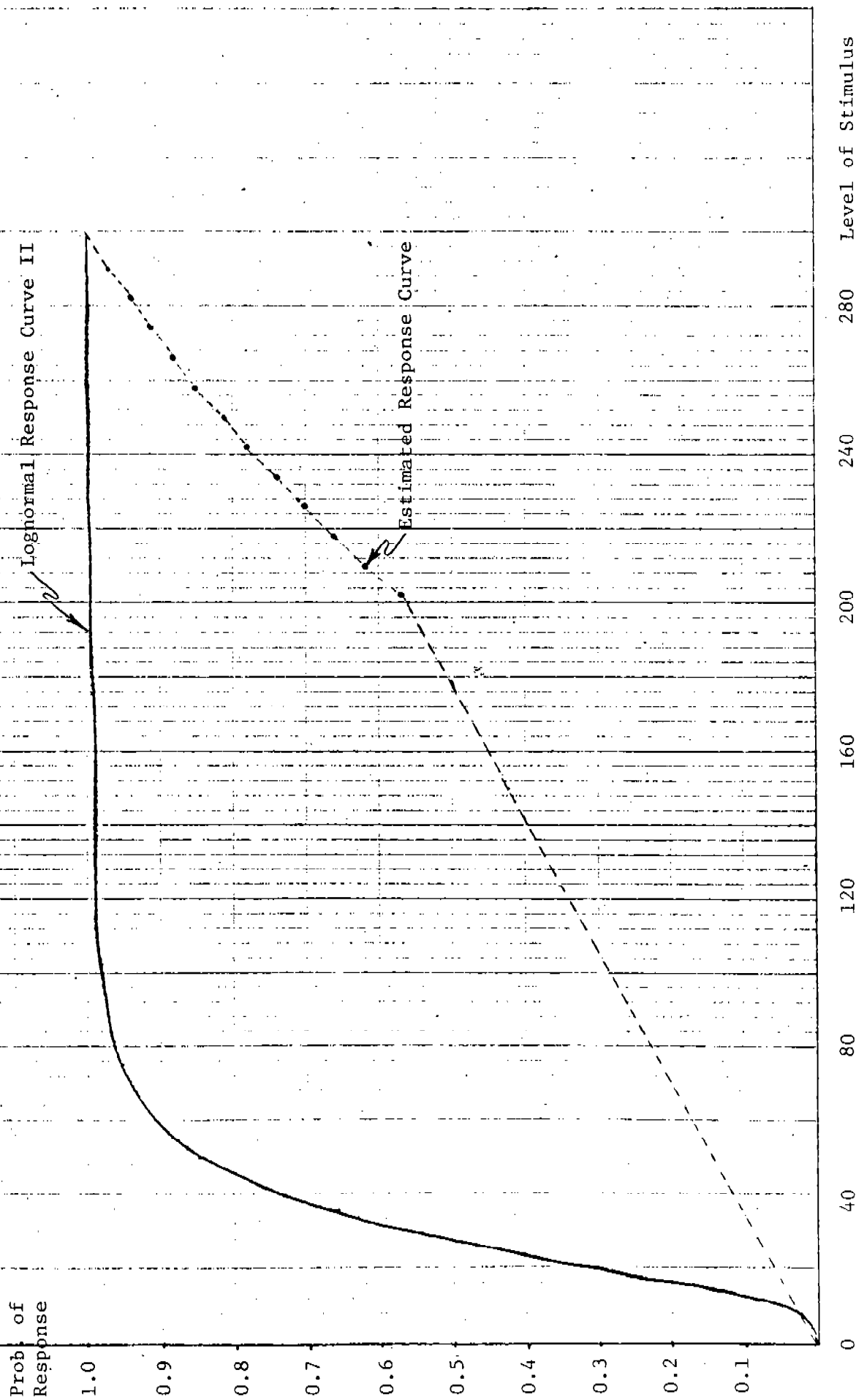


Figure 22. Design 4, lognormal II.

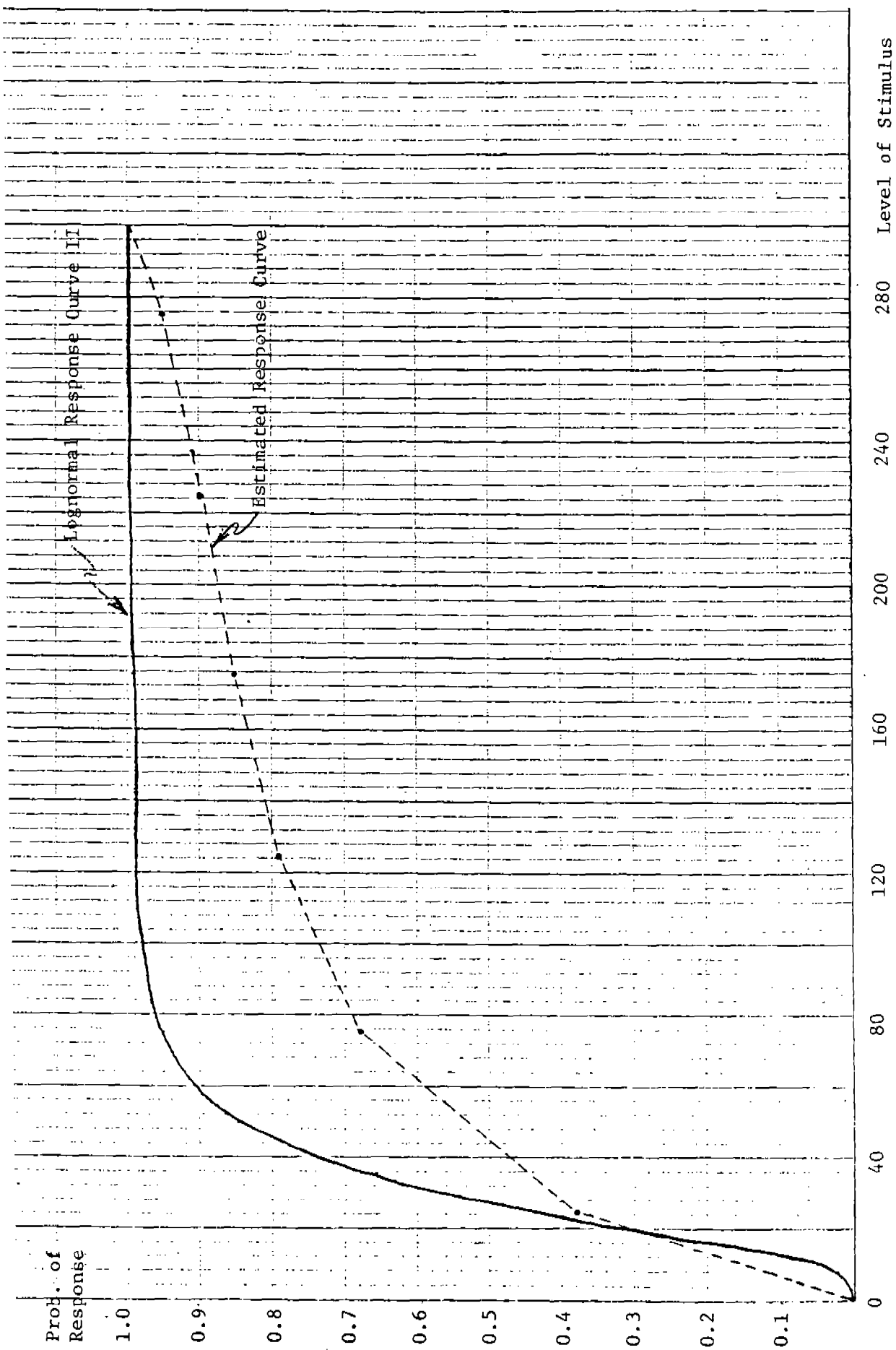


Figure 23. Sequential design, phase I, lognormal II.

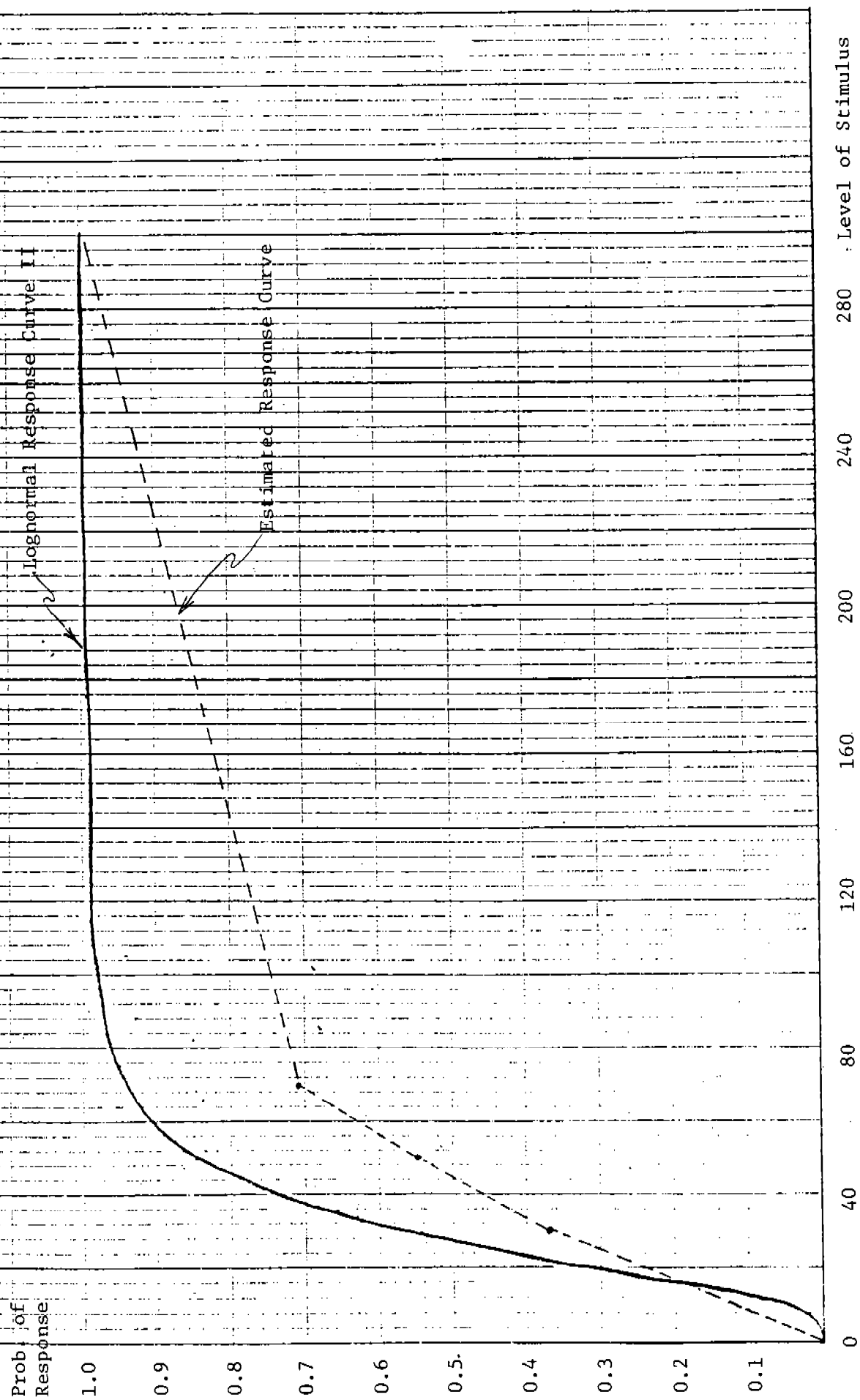


Figure 24. Sequential design, phase II, lognormal II.

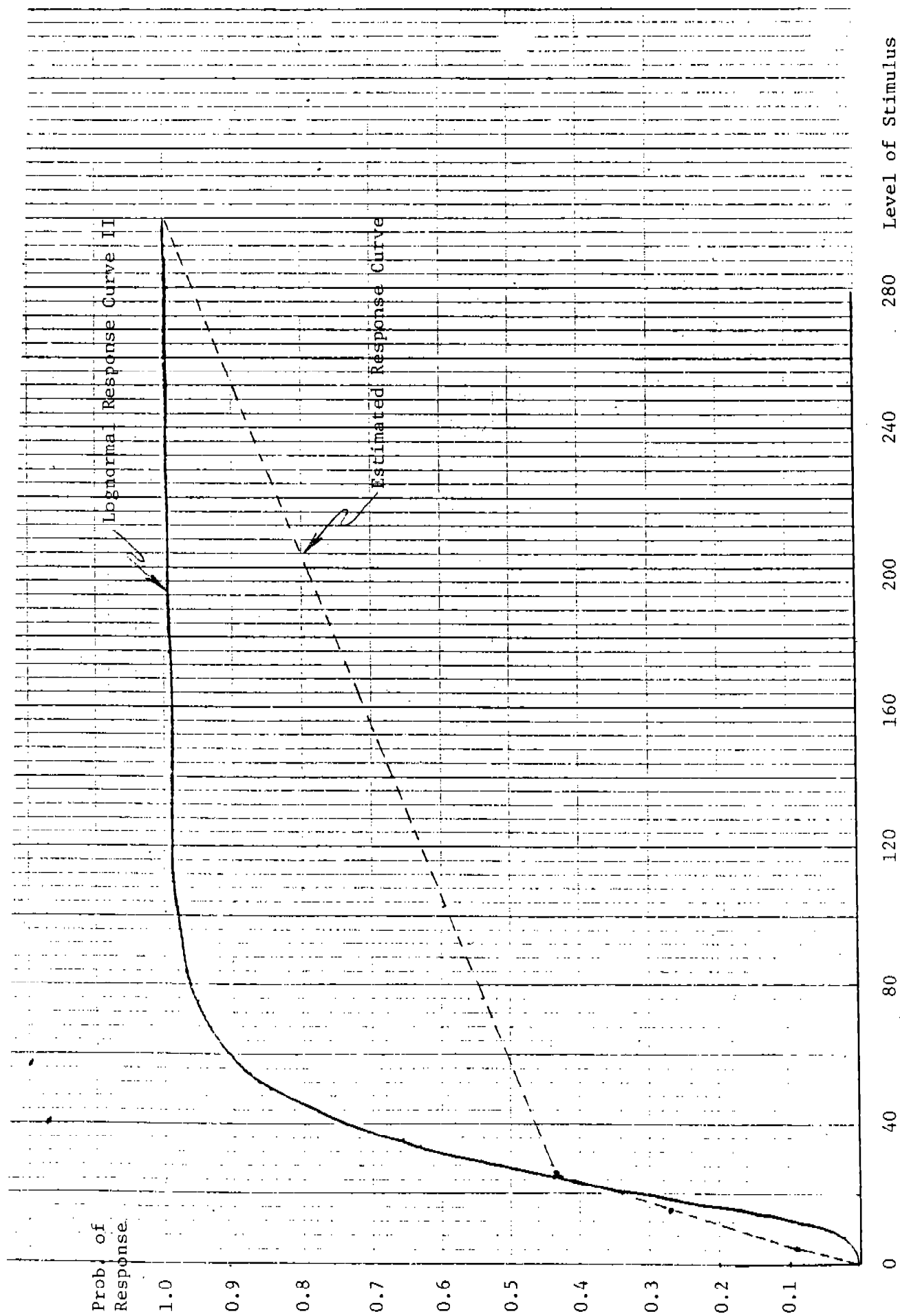


Figure 25. Sequential design, phase III, lognormal II.

Table 3. MSE's for the
Lognormal Response
Curve II

Design	MSE
1	53.7
2	55.0
3	0.4
4	144.0
5	46.8

The MSE's for the second lognormal response curve are presented in Table 3. We can observe from Table 3 that the minimum MSE is obtained for Design 3, where all 12 observations are concentrated in the center of the interval of the range of testing. The MSE obtained for Design 4 is the highest among them all. This indicates that the form of the "true" response curve affects the results significantly.

5. CONCLUSION

The application of our approach to simulated data from three types of response curve indicates that the shape of the "true" response curve is a significant factor in the evaluation of the estimate of $V_{.05}$. In real life, the "true" response curve is never known; therefore the experimenter should select his design based on his prior knowledge of the problem. Depending on the shape of the "true" response curve that is unknown to us, the $V_{.05}$ level might be underestimated or overestimated.

Sometimes the discrepancy is so large that one ends up with a high MSE for a given design, which is undesirable. The results obtained in Section 4 indicate that Design 2, where all k observations are concentrated in the left-hand half of the interval of the range of testing, has a tendency to underestimate $V_{.05}$. On the other hand, Design 4 has a tendency to overestimate $V_{.05}$. For the response curve that is considered in Section 4.3, this caused a high MSE for Design 4. The possibility of large discrepancies for these two designs makes them undesirable. The results of Section 4 also indicate that Design 3, where all k observations are concentrated in the center, gives better estimates of $V_{.05}$ in general. The discrepancies due to overestimation or underestimation are not large. This makes Design 3 more desirable than the others.

However, one should note that there is the difficulty of determining the region where the k observations will be concentrated. If the experiment has to be performed in a single phase, this region can be determined by using past information available to the investigator. Another possibility is to choose the middle portion of the interval of the range of testing.

On the basis of the analysis made, we can conclude that a design where the observations are concentrated in a region that provides the experimenter with more information is suitable for this problem. Therefore, in our analysis Design 3 is suggested as a suitable design for the estimation of $V_{.05}$. However, one should recall that the selection of the design must always be made on the basis of prior information that is available to the experimenter.

ACKNOWLEDGMENT. The work that is reported here was performed under the direction of Professor Nozer Singpurwalla. His helpful comments and advice in the preparation of the report are acknowledged.

APPENDIX

TABLES

Table A1. Data Simulated from the Weibull Response Curve

Design 1		Design 2		Design 3		Design 4		Sequential Design	
Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response
10	0	10	0	106	0	202	1	<i>Phase I:</i>	
35	0	18	0	114	1	210	1	25	0
60	1	26	0	122	0	218	1	75	1
85	0	34	0	130	1	226	1	125	1
110	1	42	0	138	1	234	1	175	1
135	1	50	1	146	1	242	1	225	1
160	1	58	1	154	1	250	1	275	1
185	1	66	1	162	1	258	1	<i>Phase II:</i>	
210	1	74	0	170	1	266	1	80	1
235	1	82	1	178	1	274	1	100	1
260	1	90	1	186	1	282	1	120	0
285	1	98	1	194	1	290	1	<i>Phase III:</i>	
								20	0
								40	0
								60	0

Table A2. Data Simulated from the Lognormal Response Curve I

Design 1		Design 2		Design 3		Design 4		Sequential Design	
Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response
10	0	10	0	106	0	202	1	<i>Phase I:</i>	
35	0	18	0	114	1	210	1	25	0
60	0	26	0	122	1	218	1	75	0
85	0	34	0	130	1	226	1	125	1
110	1	42	0	138	1	234	1	175	1
135	1	50	0	146	1	242	1	225	1
160	1	58	0	154	1	250	1	275	1
185	1	66	0	162	1	258	1	<i>Phase II:</i>	
210	1	74	0	170	1	266	1	80	0
235	1	82	0	178	1	274	1	100	1
260	1	90	1	186	1	282	1	120	1
285	1	98	1	194	1	290	1	<i>Phase III:</i>	
								20	0
								40	0
								60	0

Table A3. Data Simulated from the Lognormal Response Curve II

Design 1		Design 2		Design 3		Design 4		Sequential Design	
Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response	Level of Stimulus	Response
10	0	10	0	106	1	202	1	<i>Phase I:</i>	
35	1	18	0	114	1	210	1	25	0
60	1	26	1	122	1	218	1	75	1
85	1	34	1	130	1	226	1	125	1
110	1	42	1	138	1	234	1	175	1
135	1	50	1	146	1	242	1	225	1
160	1	58	1	154	1	250	1	275	1
185	1	66	1	162	1	258	1	<i>Phase II:</i>	
210	1	74	1	170	1	266	1	30	0
235	1	82	1	178	1	274	1	50	1
260	1	90	1	186	1	282	1	70	1
285	1	98	1	194	1	290	1	<i>Phase III:</i>	
								5	0
								15	0
								25	1

REFERENCES

- MAZZUCHI, T. A. and N. D. SINGPURWALLA (1982). The U.S. Army (BRL's) kinetic energy penetrator problem: Estimating the probability of response for a given stimulus. ARO Report 82-2, Proceedings of the Twenty-seventh Conference on the Design of Experiments in Army Research, Development, and Testing, pp. 27-58.
- MAZZUCHI, T. A. and R. SOYER (1982). Computer Programs for "A Bayesian Approach to Quantile and Response Probability Estimation Using Binary Response Data" -- A user's guide. Technical Paper Serial GWU/IRRA/TR-82/1, Institute for Reliability and Risk Analysis, The George Washington University.

INFORMATIVE QUANTILE FUNCTIONS AND IDENTIFICATION OF PROBABILITY DISTRIBUTION TYPES

Emanuel Parzen
Department of Statistics
Texas A&M University

ABSTRACT. A problem of great importance to statistical data analysts is quick identification of possible probability distributions for observed data, and classification of tail behavior of probability distributions. This paper discusses the informative quantile function $IQ(u) = \{Q(u) - Q(0.5)\} \div 2\{Q(0.75) - Q(0.25)\}$, and its use to identify probability models for observed data and its use to provide concepts of "representative distributions" which illustrate the different types of shapes and tail behavior that real distributions can have.

KEY WORDS: Quantile data analysis, informative quantile function, tail exponents, Weibull distribution, hazard function.

1. QUANTILE AND SAMPLE QUANTILE FUNCTIONS. The probability distribution of a random variable X is described in general by its distribution function $F(x) = \Pr[X \leq x]$, $-\infty < x < \infty$. When F is continuous it is described by its probability density $f(x) = F'(x)$, $-\infty < x < \infty$.

Quantile data analysis (Parzen [1979]) describes a probability distribution by

quantile function	$Q(u) = F^{-1}(u), \quad 0 \leq u \leq 1 \quad ;$
quantile density function	$q(u) = Q'(u), \quad 0 \leq u \leq 1 \quad ;$
density-quantile function	$fQ(u) = f(F^{-1}(u)) = \{q(u)\}^{-1}, \quad 0 \leq u \leq 1 \quad ;$
score function	$J(u) = -(fQ)'(u) \quad , \quad 0 \leq u \leq 1 \quad .$

Let $X_1, X_2, \dots, X_n$ be a data set. To gain insight into the processes generating the data we form the sample distribution function $F(x)$ and sample quantile function $Q(u)$. In terms of the order statistics $X_{1n} \leq X_{2n} \leq \dots \leq X_{nn}$ of the sample they are defined by

$$\begin{aligned} \tilde{F}(x) &= \frac{j}{n} \quad , \quad X_{jn} \leq x < X_{(j+1)n} \quad ; \\ \tilde{Q}(u) &= X_{jn} \quad , \quad \frac{j-1}{n} < u \leq \frac{j}{n} \quad . \end{aligned}$$

In practice we prefer to use a sample quantile function $\tilde{Q}(u)$ which is piecewise linear between the values

Research supported by the U.S. Army Research Office Grant DAAG29-83-K-0051.

$$\tilde{Q}\left(\frac{j}{n+1}\right) = X_{jn}, \quad j=1, \dots, n.$$

For graphical data analysis, we transform $\tilde{Q}(u)$ to a normalized version $IQ(u)$, called the sample informative quantile function. The value of $IQ(u)$, as u tends to 0 and 1, provide diagnostic measures of the type of probability distribution. An important classification of "type" is in terms of tail exponents (defined in section 5, but its concepts are used in the example in section 2).

2. UNITIZED AND INFORMATIVE QUANTILE FUNCTIONS. A normalization of the quantile function which depends only on its shape (and is independent of location and scale) is

$$Q_1(u) = \frac{Q(u) - \mu_1}{\sigma_1}$$

where $\mu_1 = Q(0.5)$, $\sigma_1 = Q'(0.5) = q(0.5)$. We call $Q_1(u)$ the unitized quantile function. It is the original quantile function normalized to have value 0 and slope 1 at $u = 0.5$.

One can distinguish three kinds of estimators of parameters [such as μ_1 and σ_1]: fully non-parametric [denoted $\tilde{\mu}_1$ and $\tilde{\sigma}_1$], fully parametric [denoted $\hat{\mu}_1$ and $\hat{\sigma}_1$], and functional [estimators $\hat{\mu}_1$ and $\hat{\sigma}_1$ which are the parameters of smoothed quantile functions $\tilde{Q}(u)$ obtained by smoothing the raw or fully non-parametric estimator $\tilde{Q}(u)$]. The shape of $Q(u)$ must be inferred before one can efficiently estimate μ and σ using fully parametric (or robust parametric) estimators.

A fully non-parametric estimator of $Q(0.5)$ is $\tilde{Q}(0.5)$. A fully non-parametric estimator of $q(0.5)$ is more difficult to define. We therefore consider quick and dirty approximators of $q(0.5)$ of the form

$$\sigma_p = \frac{Q(0.5 + p) - Q(0.5 - p)}{2p}$$

where $0 < p < 0.5$. We usually take $p = 0.25$; then we approximate $q(0.5)$ by

$$\sigma_{0.25} = 2\{Q(0.75) - Q(0.25)\},$$

which provides a "universal" scale parameter.

An alternative normalization to $Q_1(u)$ is

$$IQ(u) = \frac{Q(u) - Q(0.5)}{2\{Q(0.75) - Q(0.25)\}},$$

which we call the informative quantile function. It provides both graphical and numerical statistical diagnostics.

Graphically, we plot the truncated informative quantile function

$$\begin{aligned}
\text{TIQ}(u) &= -1 \text{ if } \text{IQ}(u) < -1, \\
&= 1 \text{ if } \text{IQ}(u) > 1, \\
&= \text{IQ}(u) \text{ if } |\text{IQ}(u)| \leq 1.
\end{aligned}$$

Numerically, we report the values of $\text{IQ}(u)$ at $u=0.01, 0.05, 0.10, 0.25, 0.75, 0.90, 0.95, 0.99$.

Truncating the values of $\text{IQ}(u)$ in our graphics enables us to see the "middle" of the distribution. The ends (tails) of the distributions are described numerically by the extreme values of $\text{IQ}(u)$.

For convenience in seeing at a glance in a plot of $\text{IQ}(u)$ its behavior, especially as u tends to 0 and 1, we plot on the same graph the $\text{IQ}(u)$ of a uniform distribution (it is a straight line with values -0.5 and 0.5 at $u = 0$ and 1 respectively). An empirical example is given in Section 4.

Example: Super Short Distributions. An important example of a super-short distribution ($\alpha < 0$) is $X = -\cos \pi U$ where U is uniform $[0,1]$. Since $-\cos \pi u$ is an increasing function of u , the quantile function of X is $Q(u) = -\cos \pi u$, with quantile density and density-quantile

$$q(u) = \frac{\sin \pi u}{\pi} \qquad fQ(u) = \frac{\pi}{\sin \pi u}$$

As $u \rightarrow 0$, $fQ(u) \sim u^{-1}$ so $\alpha_0 = -1$. The distribution is symmetric, in the sense that $q(1-u) = q(u)$; therefore $\alpha_1 = -1$. The interquartile range $\text{IQR} = \sqrt{2}$; the informative quantile function is $\text{IQ}(u) = (-.35) \cos \pi u$. Therefore $\text{IQ}(0) = -.35$, $\text{IQ}(1) = .35$. These values are taken as typical values of super-short distributions.

Outlying data value interpretation of $\text{IQ}(u)$. The sample informative quantile function is defined by

$$\tilde{\text{IQ}}(u) = \{\tilde{Q}(u) - \tilde{Q}(0.5)\} \div \tilde{\sigma}_1$$

where $\tilde{\sigma}_1 = 2 \text{ IQR}$ and $\text{IQR} = \tilde{Q}(0.75) - \tilde{Q}(0.25)$. The truncated sample informative quantile function $\text{TIQ}(u)$ is defined to be $\tilde{\text{IQ}}(u)$ truncated at ± 1 .

Hoaglin, Mosteller, Tukey (1983, p. 39) introduce techniques for identifying outlying (or outside) data values as those lying outside the interval

$$(\tilde{Q}(0.25) - (1.5) \text{ IQR}, \tilde{Q}(0.75) + (1.5) \text{ IQR})$$

We regard as outlying data values those lying outside the interval

$$(\tilde{Q}(0.5) - 2 \text{ IQR}, \tilde{Q}(0.5) + 2 \text{ IQR})$$

The fraction of data values which are outlying are represented on the plot of $\text{TIQ}(u)$ as values truncated to ± 1 .

3. TABLES OF TAIL VALUES OF INFORMATIVE QUANTILE FUNCTIONS. One use of the informative quantile function $\tilde{Q}(u)$ of a sample is to determine quickly probability distribution that might fit the sample. One can readily distinguish whether the data could be fit by a normal distribution or an exponential distribution [and thus determine the "probability of success" if one were to apply a more formal goodness of fit test]. However no standard parametric model may fit the data, and statistical data analysis must identify significant features of the data "non-parametrically."

Statistical scientists are seeking to define concepts which illustrate the different types of shapes and tail behavior that real distributions can have. Hoaglin, Mosteller, and Tukey (1983, p. 316) use language such as "neutral tailed (Gaussian)" and "stretch-tailed (Cauchy)." To describe the notion of tail weight, they write that it "expressed how the extreme portion of the distribution spreads out relative to the width of the center." As an index of tail behavior, they introduce (p. 323)

$$\{\tilde{Q}(0.9) - \tilde{Q}(0.1)\} \div \{\tilde{Q}(0.75) - \tilde{Q}(0.25)\} = 2\{\tilde{Q}(0.9) - \tilde{Q}(0.1)\}$$

As indices of tail behavior, this paper proposes $\tilde{Q}(u)$ at $u = 0.01, 0.05, 0.1, 0.9, 0.95, 0.99$. The true values of these indices for various familiar distributions are given in the tables. These indices are useful for exploratory data analysis of what's unusual or extraordinary about a data set, and help provide estimates of the tail exponents and tail types of distributions that might have generated the data.

Tail Values of Informative Quantile Function $\tilde{Q}(u)$

Standard Distributions

* = Approximate value of u at which $\tilde{Q}(u) = 1$.

Distribution	*	u	.01	.05	.10	.90	.95	.99
Normal	--		-.862	-.610	-.475	.475	.610	.862
Exponential	.95		-.311	-.292	-.268	.732	1.048	1.780
Logistic	.99		-1.046	-.670	-.500	.500	.670	1.046
Double Exp	.97		-1.411	-.830	-.568	.580	.830	1.411
Cauchy	.92		-7.955	-1.578	-.769	.769	1.578	7.954
Extreme Value	--		-1.346	-.828	-.599	.382	.465	0.602
Log Normal	.91		-.310	-.278	-.278	.895	1.438	3.178
Super Short	--		-.353	-.349	-.336	.336	.349	0.353

Tail Values of Informative Quantile Function IQ(u)

$$\text{Weibull } Q(u) = \{\log(1-u)^{-1}\}^{\frac{1}{\beta}}$$

* = Approximate value of u at which IQ(u) = 1.

β	*	u= .01	.05	.10	.90	.95	.99
.1	--	-1.107	-.735	-.550	.409	.505	.668
.2	--	-.921	-.655	-.506	.438	.549	.743
.3	--	-.777	-.585	-.466	.468	.595	.826
.4	--	-.662	-.525	-.430	.500	.646	.919
.5	1.0	-.571	-.473	-.396	.534	.701	1.024
.6	.98	-.498	-.427	-.366	.570	.760	1.142
.7	.97	-.437	-.387	-.338	.607	.824	1.275
.8	.96	-.388	-.351	-.312	.647	.893	1.424
.9	.95	-.346	-.320	-.295	.689	.967	1.592
1.0	.94	-.311	-.292	-.273	.732	1.048	1.780
1.1	.93	-.281	-.267	-.252	.778	1.135	1.993
1.2	.93	-.255	-.245	-.233	.827	1.229	2.232
1.3	.92	-.232	-.225	-.216	.878	1.331	2.502
1.4	.91	-.212	-.207	-.200	.931	1.440	2.806
1.5	.90	-.195	-.191	-.185	.987	1.559	3.148
1.6	.89	-.179	-.177	-.172	1.046	1.687	3.54
1.7	.89	-.165	-.163	-.159	1.107	1.825	3.969
1.8	.88	-.153	-.151	-.147	1.172	1.974	4.459
1.9	.88	-.141	-.140	-.137	1.240	2.135	5.012
2.0	.87	-.131	-.130	-.128	1.311	2.309	5.635
2.1	.87	-.121	-.121	-.119	1.386	2.497	6.338
2.2	.86	-.112	-.112	-.111	1.464	2.700	7.130
2.3	.86	-.104	-.104	-.103	1.546	2.919	8.023
2.4	.85	-.097	-.097	-.096	1.633	3.155	9.031

Tail Values of Informative Quantile Function IQ(u)

$$\text{Lognormal } Q(u) = \exp \lambda \Phi^{-1}(u)$$

* = Approximate value of u at which IQ(u) = 1.

λ	*	u= .01	.05	.10	.90	.95	.99
.5	.96	-.500	-.408	-.344	.653	.928	1.600
1	.92	-.310	-.278	-.246	.895	1.438	3.178
1.5	.88	-.203	-.192	-.179	1.223	2.260	6.655
2	.86	-.138	-.134	-.128	1.666	3.594	14.449
2.5	.84	-.096	-.094	-.092	2.266	5.761	32.083
3	.82	-.067	-.067	-.066	3.077	9.284	72.169
3.5	.81	-.048	-.047	-.047	4.175	15.012	163.511
4	.80	-.034	-.034	-.034	5.661	24.322	371.883
4.5	.80	-.024	-.024	-.024	7.673	39.454	847.538
5	.79	-.017	-.017	-.017	10.398	64.041	--
5.5	.79	-.012	-.012	-.012	14.089	103.988	--
6	.79	-.009	-.009	-.009	19.087	168.886	--
6.5	.78	-.006	-.006	-.006	25.858	274.315	--
7	.78	-.004	-.004	-.004	35.029	445.586	--
7.5	.78	-.003	-.003	-.003	47.452	723.814	--
8	.78	-.002	-.002	-.002	64.280	--	--

4. EXAMPLE OF SAMPLE INFORMATIVE QUANTILE ANALYSIS. A data set extensively discussed in a recent book on graphical methods of data analysis by Chambers, Cleveland, Kleiner, and Tukey (1983) consists of Stamford (Conn.) Monthly Maximum Ozone levels. Sample size $n=136$, sample median $\tilde{\mu}_1 = 80$, sample mean $\bar{\mu} = 89.7$, twice interquartile range $\tilde{\sigma}_1 = 147.5$, and standard deviation $\tilde{\sigma} = 52.1$. Rather than reporting the original data $X_1, \dots, X_n$ we report the normalized values $(X_j - \tilde{\mu}_1) \div \tilde{\sigma}_1$ which are used to plot $\tilde{I}\tilde{Q}(u)$; a plot of $\tilde{Q}(u)$ is given on p. 15 of Chambers et al. Numerical statistical signals are provided by the tail values:

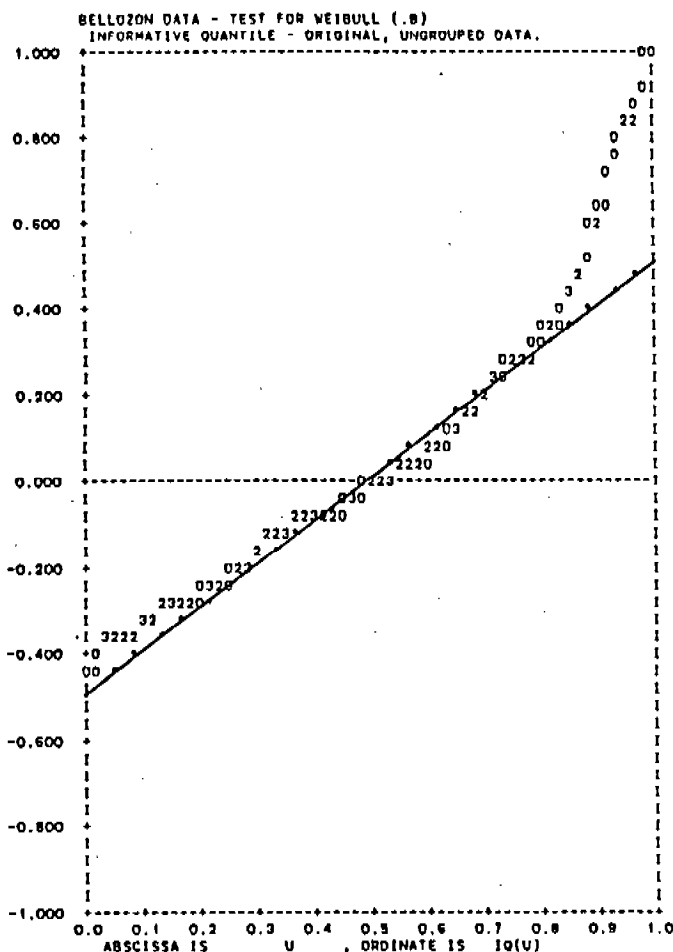
u	0.05	.1	.90	.95
$\tilde{I}\tilde{Q}(u)$	-.38	-.33	.61	.83

By consulting the table of Weibull informative quantile values, as a first guess of a distribution to fit this data one takes Weibull with parameter $\beta = 0.8$. The graph of $\tilde{I}\tilde{Q}(u)$ also suggests to us that a Weibull distribution provides a good first approximation. How to refine this approximation is a problem treated by our ONESAM data analysis program.

An alternate approach to modeling this data is to find a transformation to normality; one would then report as one's conclusion that cube root of Stamford Ozone data is normally distributed. We believe that this conclusion must be considered curve fitting, while a conclusion that the data is fit by a Weibull distribution with β in a specified range represents a curve fit with scientific insight (which may help to explain the physical mechanisms generating the data).

HELLOZON DATA - TEST FOR WEIBULL (.8)
INFORMATIVE QUANTILE - ORIGINAL, UNGROUPED DATA.

SEQUENCE WITHIN QUANTILE	ORDER STATISTICS IN QUANTERS			
	FIRST QUANTER	SECOND QUANTER	THIRD QUANTER	FOURTH QUANTER
1	-0.4475	-0.2102	0.0	0.2847
2	-0.4475	-0.1966	0.0	0.2847
3	-0.3864	-0.1898	0.0	0.2983
4	-0.3797	-0.1898	0.0	0.2983
5	-0.3797	-0.1898	0.0198	0.2983
6	-0.3797	-0.1898	0.0135	0.3051
7	-0.3797	-0.1695	0.0203	0.3051
8	-0.3661	-0.1424	0.0339	0.3458
9	-0.3593	-0.1356	0.0407	0.3593
10	-0.3525	-0.1288	0.0407	0.3661
11	-0.3525	-0.1288	0.0475	0.3797
12	-0.3525	-0.1085	0.0475	0.4136
13	-0.3322	-0.1085	0.0475	0.4203
14	-0.3322	-0.1085	0.0610	0.4271
15	-0.3254	-0.1085	0.0746	0.4475
16	-0.3254	-0.0949	0.0814	0.4746
17	-0.3186	-0.0949	0.0949	0.4881
18	-0.2915	-0.0814	0.0949	0.5085
19	-0.2847	-0.0814	0.1220	0.6034
20	-0.2847	-0.0814	0.1288	0.6034
21	-0.2847	-0.0746	0.1288	0.6102
22	-0.2847	-0.0610	0.1356	0.6305
23	-0.2847	-0.0610	0.1424	0.6273
24	-0.2847	-0.0610	0.1559	0.7322
25	-0.2847	-0.0610	0.1559	0.7593
26	-0.2847	-0.0610	0.1559	0.7864
27	-0.2712	-0.0610	0.1898	0.8203
28	-0.2576	-0.0542	0.2102	0.8271
29	-0.2508	-0.0542	0.2237	0.8271
30	-0.2305	-0.0475	0.2237	0.8542
31	-0.2237	-0.0339	0.2309	0.8949
32	-0.2237	-0.0339	0.2576	0.9153
33	-0.2237	0.0	0.2644	1.0169
34	-0.2237	0.0	0.2644	1.0847



Printer
Plot of
Truncated
Sample
Informative
Quantile
Function
at Stamford
Monthly
Maximum
Ozone
Levels,
n=136

5. TAIL EXPONENTS CLASSIFICATION OF PROBABILITY LAWS. From extreme value theory, statisticians have long realized that it is useful to classify distributions according to their tail behavior (behavior of $F(x)$ as x tends to $+\infty$). It is usual to distinguish three main types of distributions, called (1) limited, (2) exponential, and (3) algebraic. This classification can also be expressed in terms of the density quantile function $fQ(u)$; we call the types short, medium, and long tail.

A reasonable assumption about the distributions that occur in practice is that their density-quantile functions are regularly varying in the sense that there exist tail exponents α_0 and α_1 such that, as $u \rightarrow 0$,

$$fQ(u) = u^{\alpha_0} L_0(u) \quad , \quad fQ(1-u) = u^{\alpha_1} L_1(u)$$

where $L_j(u)$ for $j=0,1$ is a slowly varying function.

A function $L(u)$, $0 < u < 1$ is usually defined to be slowly varying as $u \rightarrow 0$ if, for every y in $0 < y < 1$, $L(yu)/L(u) \rightarrow 1$ or $\log L(yu) - \log L(u) \rightarrow 0$. For estimation of tail exponents we will require further that, as $u \rightarrow 0$,

$$\int_0^1 \{ \log L(yu) - \log L(u) \} dy \rightarrow 0$$

which we call integrally slowly varying. An example of a slowly varying function is $L(u) = \{\log u^{-1}\}^\beta$.

Classification of tail behavior of probability laws. A probability law has a left tail type and a right tail type depending on the value of α_0 and α_1 . If α is the tail exponent, we define:

$\alpha < 0$	super short tail
$0 \leq \alpha < 1$	short tail
$\alpha = 1$	medium tail
$\alpha > 1$	long tail

Medium tailed distributions are further classified by the value of $J^* = \lim J(u)$:

$\alpha = 1$, $J^* = 0$	medium long tail
$\alpha = 1$, $0 < J^* < \infty$	medium-medium tail
$\alpha = 1$, $J^* = \infty$	medium-short tail

One immediate insight into the meaning of tail behavior is provided by the hazard function

$$h(x) = f(x) \div \{1-F(x)\}$$

with hazard quantile function $hQ(u) = fQ(u) \div 1-u$. The convergence behavior of $h(x)$ as $x \rightarrow \infty$ is the same as that of $hQ(u)$ as $u \rightarrow 1$. From the definitions one sees that $h^* = \lim_{x \rightarrow \infty} h(x)$ satisfies

$h^* = \infty$	(increasing hazard rate)	Short or medium-short tail
$0 < h^* < \infty$	(constant hazard rate)	Medium-medium tail
$h^* = 0$	(decreasing hazard rate)	Long or medium-long tail

Formulas for computing tail exponents. The representation of $fQ(u)$ suggests a formula for computation of tail exponents α_0 and α_1 (which may be adapted to provide estimators from data).

Theorem: Computation of tail exponents

$$-\alpha_0 = \lim_{u \rightarrow 0} \int_0^1 \{\log fQ(yu) - \log fQ(u)\} dy$$

Equivalently

$$-\alpha_0 = \lim_{p \rightarrow 0} \frac{1}{p} \int_0^p \log fQ(t) dt - \log fQ(p)$$

Similarly

$$\begin{aligned} \alpha_1 &= \lim_{u \rightarrow 0} \int_0^1 \{\log fQ(1-yu) - \log fQ(1-u)\} dy \\ &= \lim_{p \rightarrow 1} \frac{1}{1-p} \int_p^1 \log fQ(t) dt - \log fQ(1-p) \end{aligned}$$

Proof: $\log fQ(u) = \alpha_0 \log u + \log L_0(u),$

$$\log fQ(yu) - \log fQ(u) = \alpha_0 \log y + \log L_0(yu) - \log L_0(u).$$

Since $\int_0^1 \log y dy = -1$, we conclude that

$$\int_0^1 \{\log fQ(yu) - \log fQ(u)\} dy = -\alpha_0 + o(u)$$

Similarly one derives formula for α_1 .

Because the density-quantile and quantile-density functions are reciprocals, we obtain similar formulas for $q(u)$ which may be easier to implement in practice:

$$q(u) = u^{-\alpha_0} L_0(u), \quad \text{as } u \rightarrow 0,$$

$$q(u) = (1-u)^{-\alpha_1} L_1(1-u), \quad \text{as } u \rightarrow 1;$$

$$\alpha_0 = \lim_{u \rightarrow 0} \int_0^1 \{\log q(yu) - \log q(u)\} dy;$$

$$\alpha_1 = \lim_{u \rightarrow 0} \int_0^1 \{\log q(1-yu) - \log q(1-u)\} dy.$$

Practical implementation of the foregoing estimators of tail exponents remains to be investigated. Related estimators are given in Mason (1982) and the papers referenced there.

REFERENCES

- Chambers, J. M., Cleveland, W. S., Kleiner, B., Tukey, P. A. (1983) Graphical Methods for Data Analysis, Duxbury: Boston.
- Hoaglin, C., Mosteller, F. and Tukey, J. W. (1983) Understanding Robust and Exploratory Data Analysis, Wiley: New York.
- Mason, D. M. (1982) Laws of large numbers for extreme values. Annals of Probability, 10, 754-764.
- Parzen, E. (1979) Nonparametric Statistical Data Modeling. Journal of the American Statistical Association, 74, 105-131.

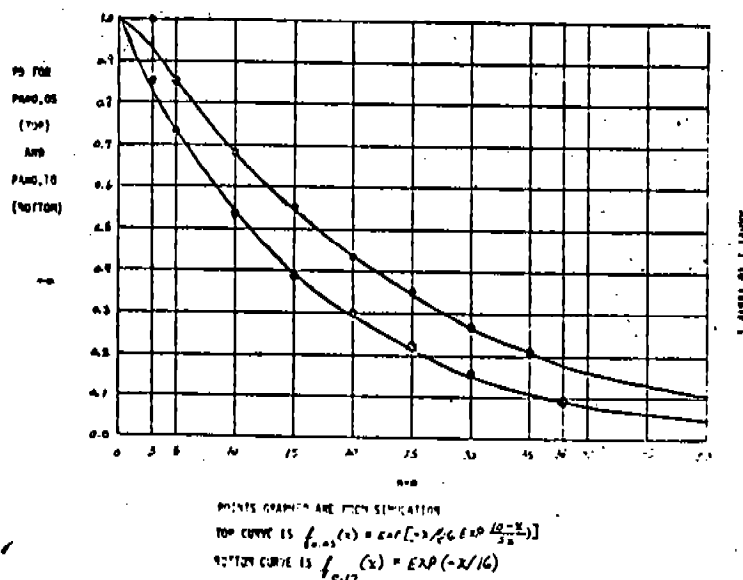
ON THE LEHMANN POWER ANALYSIS FOR THE WILCOXON RANK SUM TEST

James R. Knaub, Jr.

US Army Logistics Center

ABSTRACT

The Wilcoxon Rank Sum (or Mann-Whitney) Test is among the most useful and powerful of the non-parametric hypothesis tests. However, as with many hypothesis tests, when a clear alternative hypothesis and corresponding power analysis is not present, the practical interpretation of results using this test suffers greatly. This paper presents and clarifies an alternative suggested by E. L. Lehmann in 1953 and provides tables of practical use which have not previously been calculated due to computational difficulties.



On the Lehmann Power Analysis for the Wilcoxon Rank Sum Test

The Wilcoxon Rank Sum (or Mann-Whitney) Test is among the most useful and powerful of the non-parametric hypothesis tests. However, as with many hypothesis tests, when a clear alternative hypothesis and corresponding power analysis is not present, the practical interpretation of results using this test suffers greatly. This paper presents and clarifies an alternative suggested by E. L. Lehmann in 1953 (Annals of Mathematical Statistics [7]) and provides tables of practical use which have not previously been calculated due to computational difficulties. This work has recently been applied to survey data gathered for the US Army Logistics Center. (See reference [5].)

When sample sizes are small, and a power analysis is not available, one may fail to reject the null hypothesis when the true state of nature is very different from what is stated in the null hypothesis. With a small sample size and small α , it may be impossible to reject H_0 . Further, when sample sizes are very large, the null hypothesis may be rejected at a very small significance level when actually the null hypothesis is so nearly true, that it is close enough for all practical purposes. Taken to the extreme, with infinite sample sizes, the attained significance level will be zero, even when there is only a very small, but finite difference between H_0 and the true state of nature. Thus significance level can be very misleading if used alone.

When a null and a definitive alternative hypothesis can both be stated, and probability distributions found under each, the results of an hypothesis test can be stated similarly to a confidence interval if the "point estimate" from the observed values falls between the two hypotheses. In the case of the Wilcoxon Rank Sum Test, only one alternative hypothesis has been well developed and will be presented here. Due to the nature of this test, however, even if the evidence may strongly indicate that the true state of nature is not bounded between this alternative and the null hypothesis, this power analysis can still be used to obtain a reasonable estimate of what the actual state of nature happens to be. (In the case of the Multiple-sample Westenberg-type tests of reference [4], an alternative must be picked such that the true state of nature is indicated to be bounded by the null and alternative hypotheses. Fortunately, that is not the case here, nor was it the case in reference [6], which is a multi-sample test.)

Consider that the null hypothesis, H_0 , of the Wilcoxon Rank Sum Test indicates that $P(X < Y) = 1/2$. That is, under H_0 , any value picked at random from the Y population, is larger than any value picked at random from the X population, with probability of 1/2. Here an alternative hypothesis, H_1 , is used such that $P(X < Y) = 2/3$. (The exact form of H_1 is discussed in [7].)

Graph 1 illustrates a possible configuration for this alternative hypothesis. For this example, consider that under H_0 , all observations are taken from a $N(r,s)$ distribution such as the $N(5,1)$ shown on the left in graph 1, but under H_1 , the Y sample comes from the $N(r+0.61s, s)$ distribution, while the X sample comes from the $N(r,s)$ distribution.

Another example of a possible situation satisfying the alternative hypothesis, H_1 , given approximately by comparing a gamma (4,1) with a gamma (3,1), is illustrated by graph 2.

Note that the Wilcoxon Rank Sum Test is most sensitive to location, a little sensitive to shape, but not to dispersion (except as it relates proportionately to differences in location). Therefore, it is the differences in location that are of primary importance in graphs 1 and 2.

In order to determine the probability of drawing a value from distribution A which is larger than a simultaneously drawn value from distribution B, the following may be used:

$$P = \int_{x=-\infty}^{\infty} f_B(x) \int_{t=x}^{\infty} f_A(t) dt dx$$

where f_A and f_B represent density functions.

For the case where A and B are both gamma distributions,

$$P = 1 - \frac{\beta_A^{-\alpha_A} \beta_B^{-\alpha_B}}{\beta_A^{\alpha_A} \beta_B^{\alpha_B}} \frac{1}{(\alpha_A - 1)!} \frac{\alpha_B - 1}{r!} \frac{\beta_B^{r+1}}{(\alpha_B - 1 - r)!} \frac{(\alpha_A + \alpha_B - 2 - r)!}{([\alpha_A + \beta_B] / \beta_A \beta_B)^{\alpha_A + \alpha_B - 1 - r}}$$

For gamma (4,1) and gamma (3,1), $P = 21/32 \approx 0.656$.

For normal distributions, use $\Phi[(\mu_A - \mu_B)/\sqrt{\sigma_A^2 + \sigma_B^2}]$, as in the Church-Harris-Downton (C-H-D) method of missile motor safety testing [2]. (Note: This reference to the C-H-D method should not be construed as the author's endorsement of this method for the purpose of missile motor safety testing.)

The calculation of power under this alternative involves a summation over a typically large number of products. Calculation of this value can become extremely time consuming, even for a high speed computer. A program was written for the author at White Sands Missile Range which will calculate these exact values, however, in general, the sample sizes must be very small. Recently, however, the author constructed a simulation which provides estimates of the power for much larger sample sizes. A number of the "products" mentioned earlier are calculated and the mean is computed. The number of products involved in the exact calculation can be determined, and it is multiplied by this mean. Comparison to values calculated exactly (when practical), and a study of the sensitivity of the results to increased replications, as well as comparison to other simulated values bounding the results in the tables, led to the use of from 1 to 20 million replications to simulate values for the tables found in this paper. (Work has been done, reference [3], to determine the number of simulation replications needed under less radical circumstances. Here, however, a larger number of replications appears necessary.) (For $n = m$

= 50, up to 35 million replications were used. It appeared, however, that fewer replications using a number of different seeds yielded mean answers which more quickly converged to reasonable results, especially when using antithetic seeds.)

In the tables, n is the sample size of the X sample, m is the sample size of the Y sample, RS is the rank sum for which type I and type II error probabilities are calculated, PA is the former of those probabilities, and PB is the later. Specifically, PA is the attained probability of making an error if H_0 is rejected, and PB is the attained probability of error if H_1 is rejected, both corresponding to the same RS value. RS is always calculated by adding the ranks of the Y elements in the combined sample. Note that for smaller sample sizes, power + PB is noticeably larger than unity due to the discrete nature of this test. That is, the probability of obtaining exactly the event observed (and no other) is non-zero.

Three significant digits are given for PA and only two for power and PB simply because it takes fewer replications of the simulation to satisfactorily obtain a value for PA than for the others.

From the annex to table 1, it is found empirically that if x is the size of each of the two samples, and $f_\alpha(x)$ is the probability of a type II error under the alternative used here, adjusted to correspond to a specific significance level, then, as a continuous representation of actually a discrete process,

$$f_{0.10}(x) \approx \exp(-x/16)$$

for at least $3 \leq x \leq 40$, and perhaps this approximation could be trusted for $x = 45$ or larger. However, extrapolations are always more dangerous than interpolations, so caution is advised for further extensions.

For $\alpha = 0.05$,

$$f_{0.05}(x) \approx \exp(-x/[26 \exp \frac{10-x}{5x}])$$

for at least $4 \leq x \leq 40$, and perhaps for x substantially larger. Using this approximation, it is conjectured that for $n = m = 66$, when PA is approximately 0.05 ($RS = 4751$), then PB for this alternative is also approximately 0.05 and the true state of nature would then quite safely be said to (probably) lie between the null and alternative hypotheses. (At the 0.1 probability level for PA and PB , this could be said when $n = m = 37$, and $RS = 1507$.) An extrapolation to $n = m = 66$ is questionable, however, and further extrapolation is not advised. Computer simulation for $n = m = 50$ indicates that for the top curve ($PA \approx 0.05$) in Annex I to table 1, true values in this area for PB may be somewhat smaller than this curve predicts. For $PA \approx 0.10$, PB values for large n and m may be somewhat larger than predicted.

In Conover's book [1], an approximation is given to find RS for a given PA value. ($RS \approx m(m+n+1)/2 + x_{1-\alpha} \sqrt{mn(n+m+1)/12}$, where $x_{1-\alpha}$ is from the table of the cumulative normal distribution.) The two functions given earlier can be used to estimate PB values when $PA \approx 0.10$ or 0.05 .

The final graphs, 3-7, are taken from work the author directed at White Sands Missile Range in order to study this alternative for the Wilcoxon Rank Sum Test with emphasis on simulation validation for missile flight simulations. When comparing a very few live firings to a substantially larger number of simulations for each scenario, it can be seen from these graphs that once one sample is substantially larger than the other, increasing the larger sample size further does very little to improve the power. These graphs are continuous representations of what are actually discrete points. The values for those points were calculated analytically as noted in the acknowledgements.

Finally, when $n \neq m$, PB can be bounded using the exponential formulations found earlier in this paper. If, for example, RS is such that $PA = 0.1$, and x_1 is the smaller of n and m , and x_2 is the larger, then one has that approximately $\exp(-x_2/16) < PB < \exp(-x_1)$, with PB somewhat closer to $\exp(-x_1/16)$, especially when $x_1 \ll x_2$.

For larger sample sizes than are handled here, parametric methods may be used. However, in addition to the probability of error associated with any conclusion drawn from a parametric test, there is the additional risk involved in assuming the distributional forms used in such a test. Hypothesis tests should also be used to study these distributional assumptions to provide a more complete risk analysis.

EXAMPLE:

Consider two sources of data, X and Y, where it is suspected that Y may represent a population of larger location than X, but this is not clear. If 11 observations are taken from the X population, and 19 observations taken from Y, then the critical value of the rank sum (RS) of the Y sample observations within the combined sample which represents the point at which rejection of the null hypothesis would occur using $\alpha = 0.10$, is approximately

$$\begin{aligned} RS &\approx m(m+n+1)/2 + 1.2816\sqrt{mn(m+n+1)/12} \\ &= (19)(31)/2 + 1.2816\sqrt{(19)(11)(31)/12} \\ &\approx 324.3 \end{aligned}$$

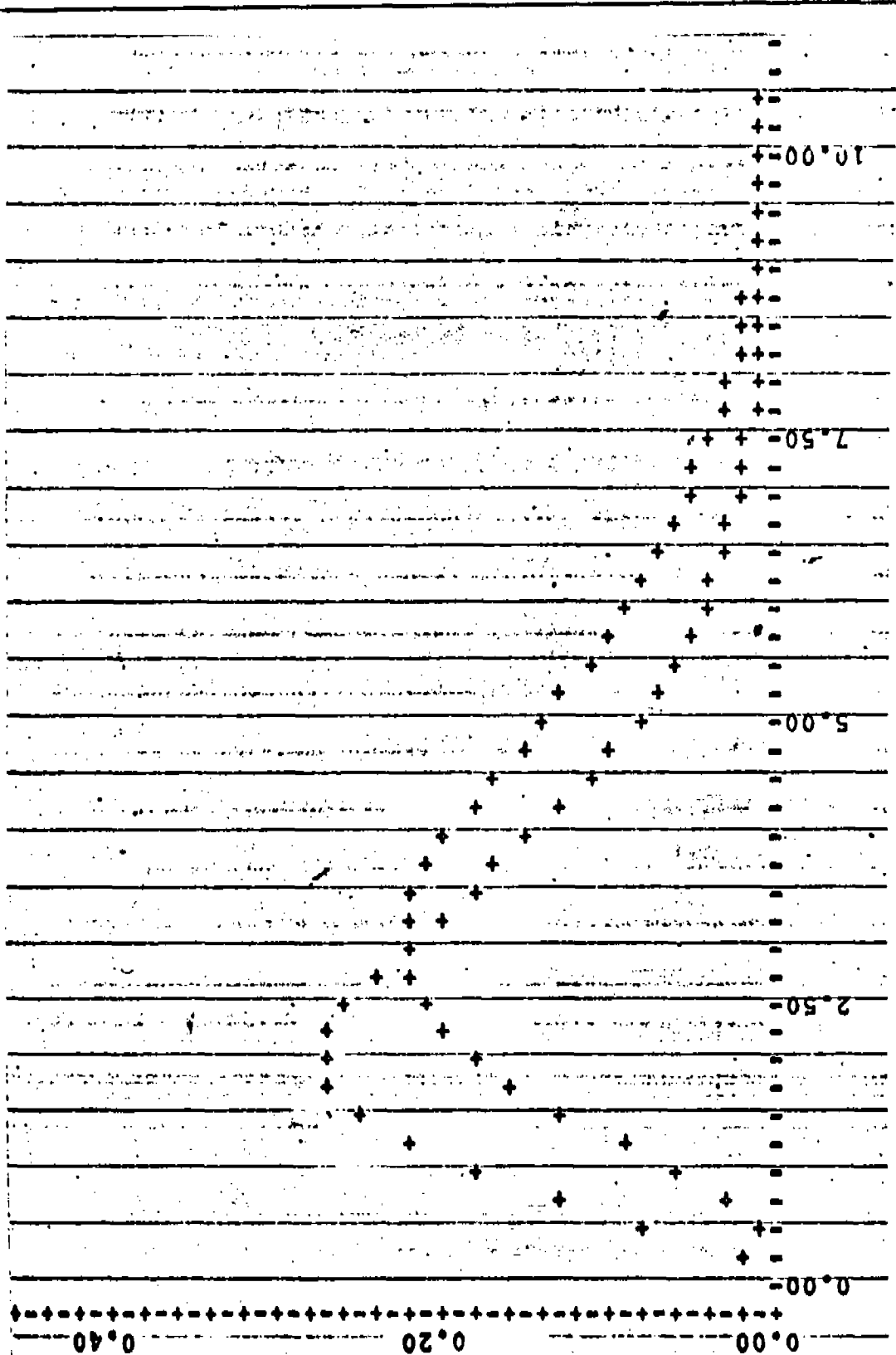
Therefore, if $RS \geq 325$, H_0 would be rejected at the $\alpha = 0.10$ level. However, should $RS = 325$, and H_0 not be rejected, then the probability of making a type II error with respect to the alternative hypothesis illustrated in graphs 1 and

2 is approximately bounded by $\exp(-19/16)$ and $\exp(-11/16)$, so $0.30 < PB < 0.50$. Note that, from table 2, when $PA = 0.099$, $PB(10,20) \approx 0.43$. Using 4,000,000 replications in the program given in Appendix A, for $m = 19$, $n = 11$, and $RS = 325$, resulted in $PA = 0.100$ and $PB = 0.42$.



Alternative Hypothesis (H_1) Using a
 $N(5,1)$ and an $N(5.61,1)$

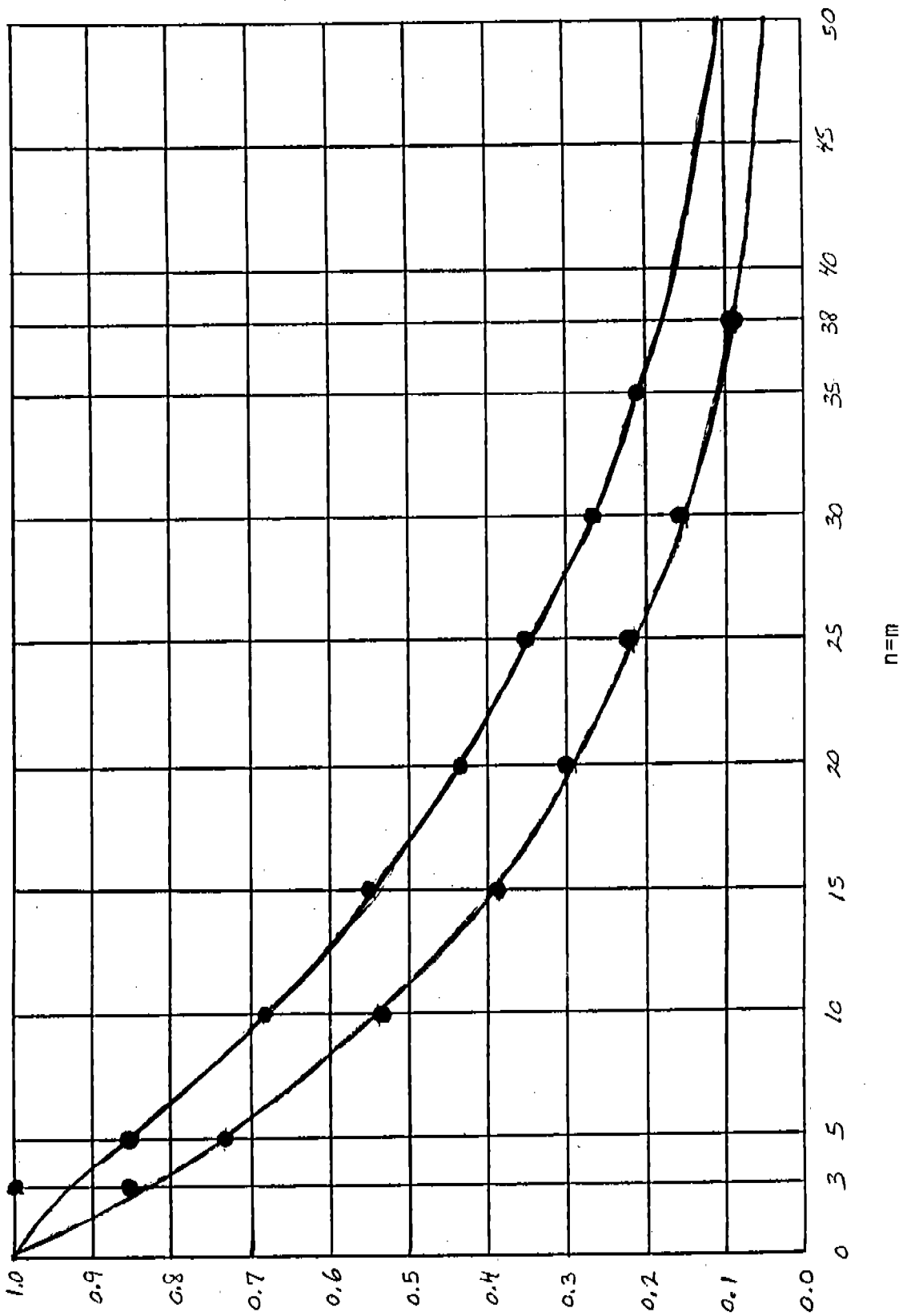
Graph 1



Graph 2

Table 1

n = m	RS	PA	power	PB
3	12	0.350	0.62	0.56
3	14	0.100	0.27	0.85
3	15	0.050	0.15	1.00
5	32	0.210	0.54	0.55
5	34	0.111	0.37	0.71
5	35	0.075	0.29	0.79
5	36	0.048	0.21	0.86
5	39	0.008	0.05	0.97
10	122	0.108	0.52	0.51
10	123	0.095	0.49	0.54
10	127	0.052	0.36	0.67
10	128	0.045	0.34	0.69
10	136	0.009	0.13	0.89
15	264	0.101	0.63	0.39
15	265	0.094	0.61	0.41
15	273	0.049	0.47	0.55
15	289	0.009	0.21	0.80
20	458	0.101	0.71	0.30
20	459	0.096	0.70	0.30
20	471	0.051	0.58	0.43
20	472	0.048	0.57	0.44
20	496	0.010	0.30	0.71
25	704	0.101	0.79	0.22
25	705	0.098	0.78	0.23
25	723	0.050	0.66	0.35
25	758	0.009	0.38	0.63
30	1002	0.101	0.85	0.16
30	1003	0.099	0.84	0.16
30	1027	0.050	0.74	0.27
30	1073	0.010	0.47	0.54
35	1383	0.050	0.79	0.21
38	1587	0.100	0.91	0.09



POINTS GRAPHED ARE FROM SIMULATION

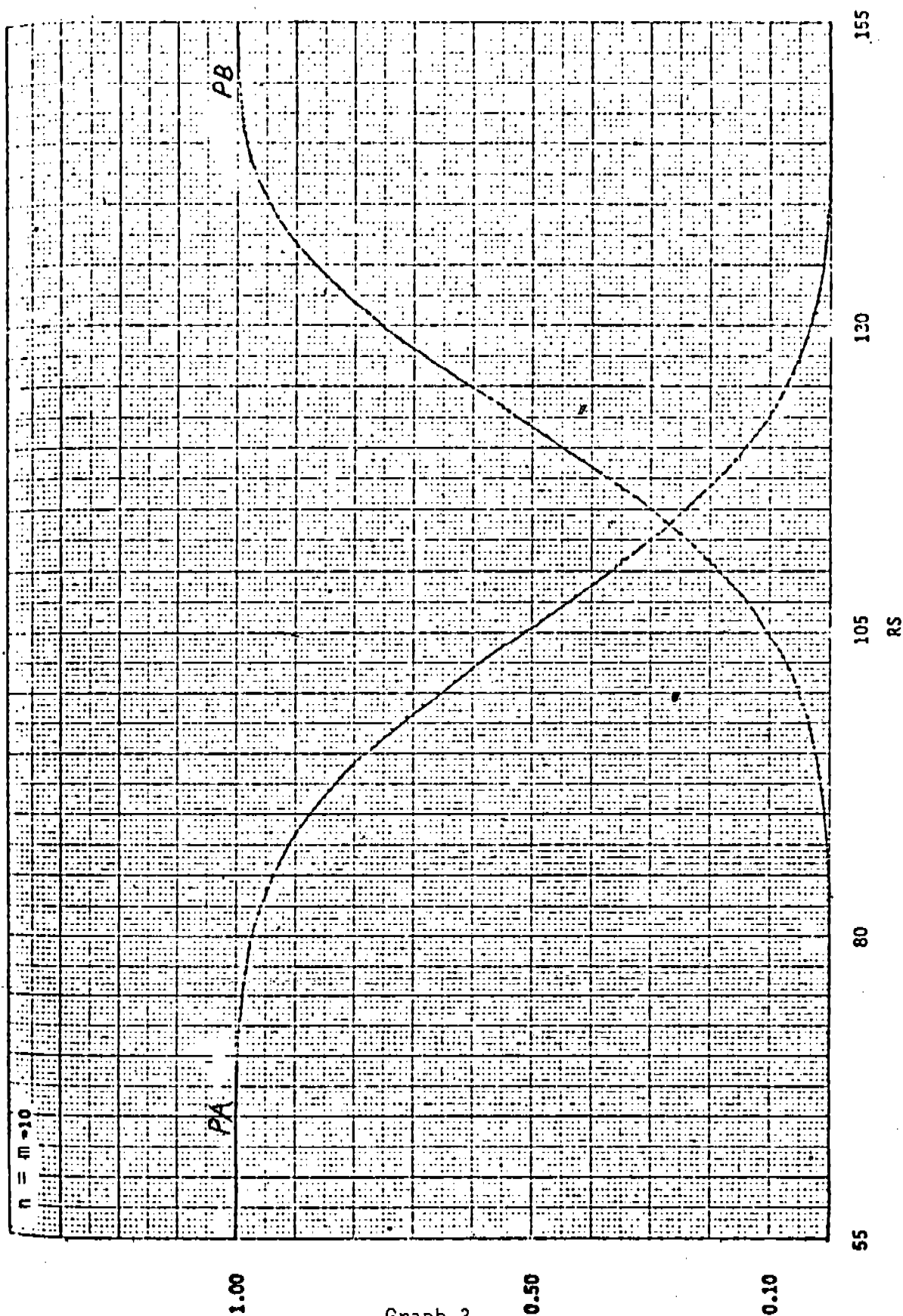
TOP CURVE IS $f_{0.05}(x) = \exp\left[-x / \left(26 \exp \frac{10 - x}{5x}\right)\right]$

BOTTOM CURVE IS $f_{0.10}(x) = \exp(-x/16)$

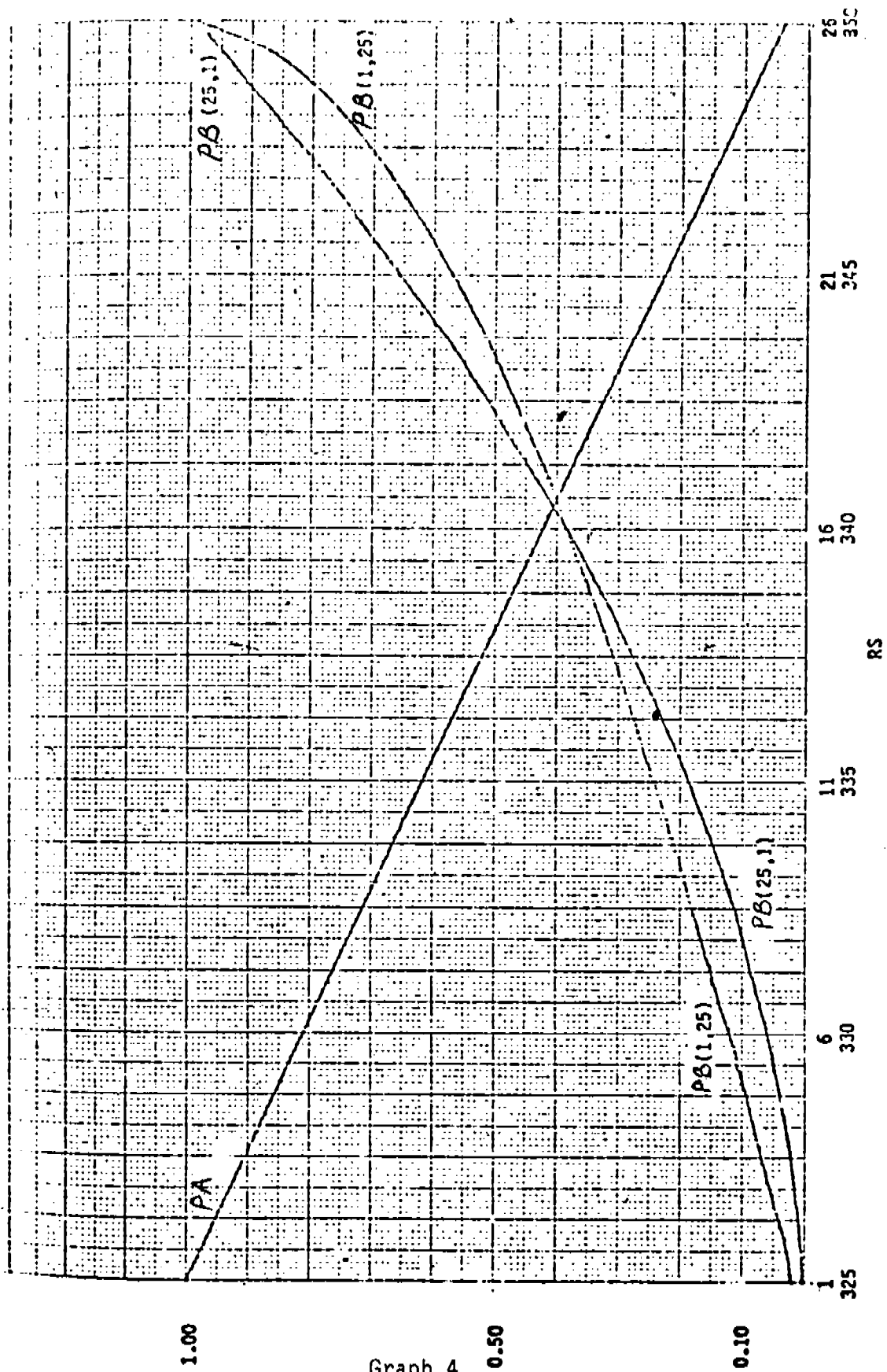
3 FOR
A=0.05
(TOP)
AND
A=0.10
(BOTTOM)

Table 2

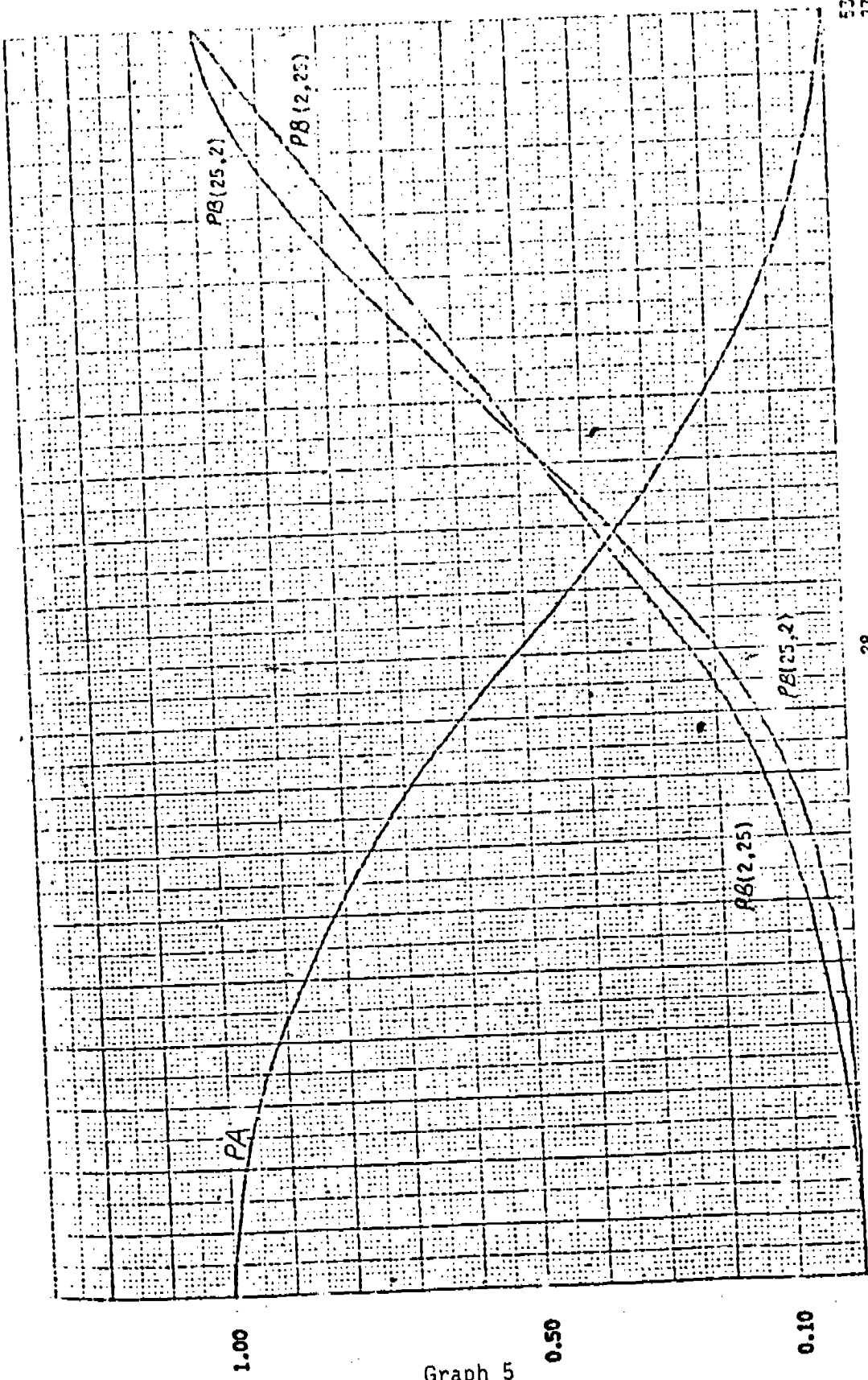
$\begin{matrix} n, m \\ n \neq m \end{matrix}$	RS	PA	power	PB
5,10	85	0.297	0.70	0.35
5,10	91	0.103	0.42	0.63
5,10	92	0.082	0.37	0.68
5,10	94	0.050	0.27	0.77
5,10	99	0.010	0.10	0.93
10,5	45	0.297	0.71	0.34
10,5	51	0.103	0.41	0.65
10,5	52	0.082	0.35	0.70
10,5	54	0.050	0.26	0.79
10,5	59	0.010	0.08	0.95
5,25	412	0.094	0.45	0.57
5,25	418	0.048	0.33	0.69
5,25	429	0.009	0.13	0.88
5,25	430	0.008	0.12	0.89
25,5	102	0.094	0.44	0.59
25,5	108	0.048	0.29	0.73
25,5	119	0.009	0.09	0.92
25,5	120	0.008	0.08	0.93
10,20	340	0.099	0.58	0.43
10,20	348	0.050	0.44	0.58
10,20	363	0.009	0.20	0.81
20,10	185	0.099	0.59	0.43
20,10	193	0.050	0.44	0.58
20,10	208	0.010	0.18	0.84
5,50	1444	0.105	0.50	0.51
5,50	1457	0.050	0.36	0.65
5,50	1480	0.008	0.14	0.87
50,5	184	0.105	0.50	0.52
50,5	197	0.050	0.32	0.69
50,5	220	0.008	0.09	0.92
10,50	1590	0.101	0.65	0.36
10,50	1608	0.051	0.52	0.49
10,50	1643	0.009	0.26	0.75
50,10	370	0.102	0.68	0.33
50,10	388	0.051	0.52	0.49
50,10	423	0.009	0.22	0.79



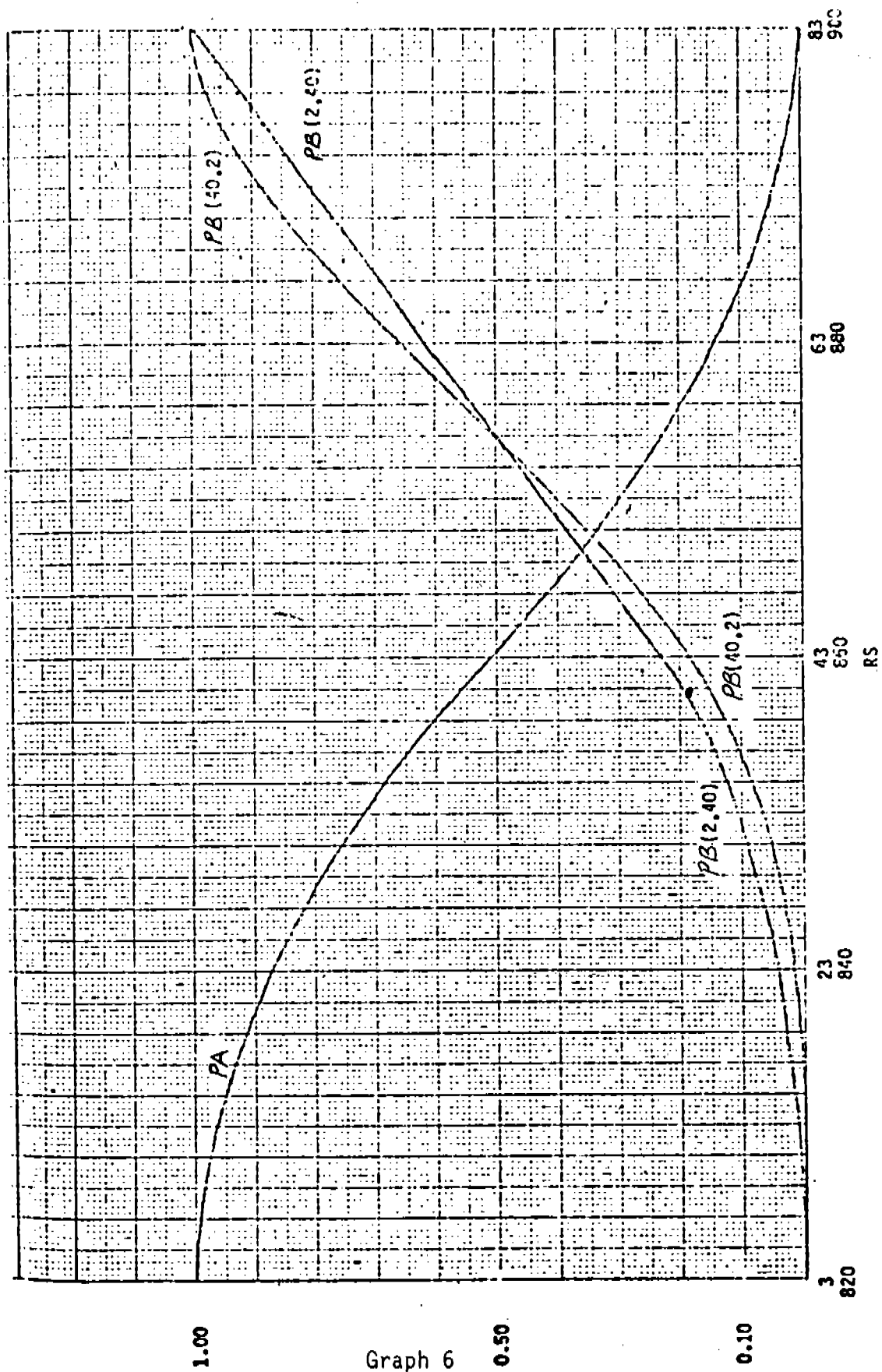
Graph 3



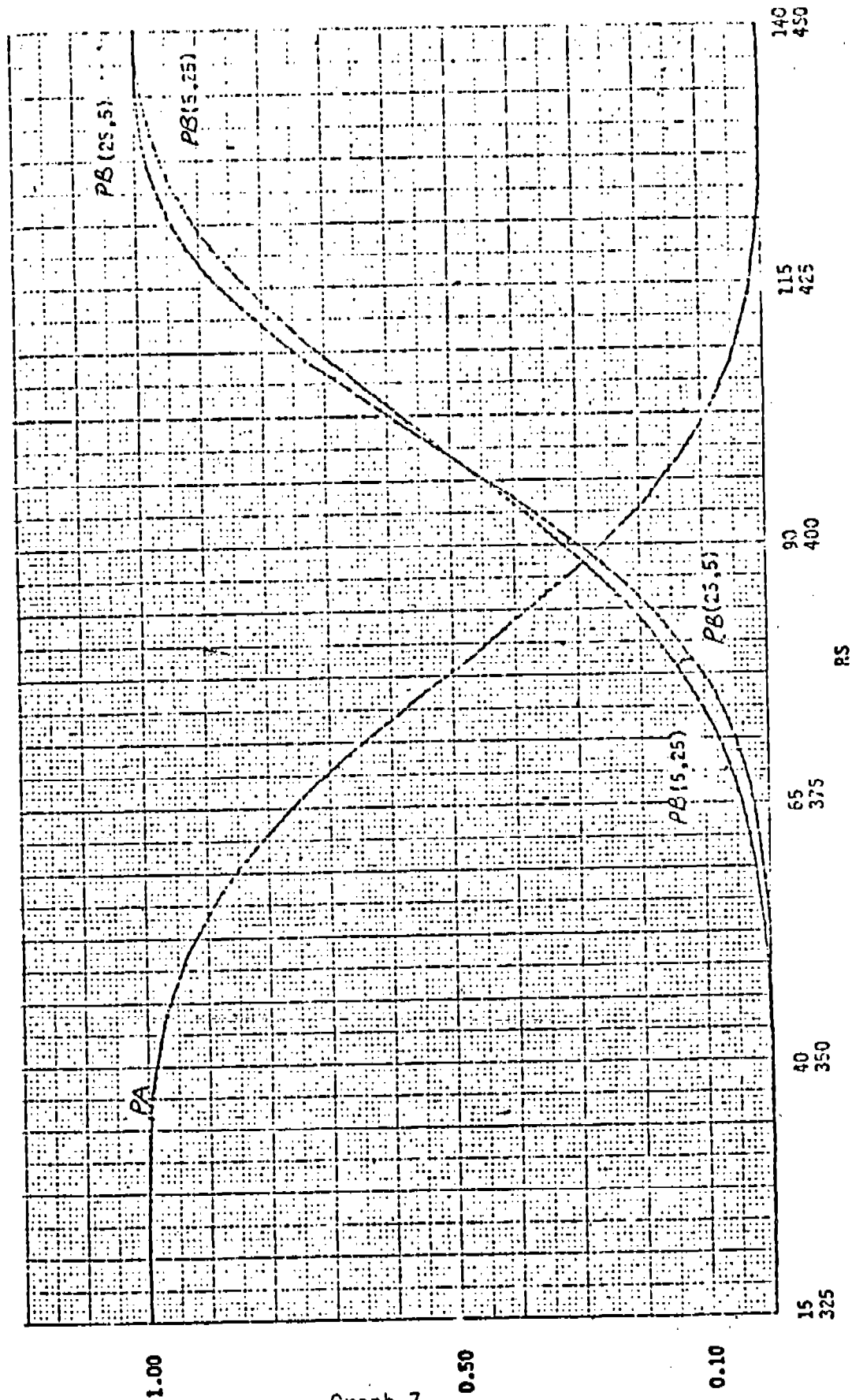
Graph 4



Graph 5



Graph 6



Graph 7

APPENDIX A
FORTRAN CODE FOR
SIMULATION:
"LEHMANN POWER ANALYSIS
FOR THE
WILCOXON RANK SUM TEST"
(LPAWRST)

VAX/VMS	KNAUB	LPAWRST 14-JUL-1983 15:35	LPA01 14-JUL-1983 15:35	USER:[KNAUB]LPAWRST
VAX/VMS	KNAUB	LPAWRST 14-JUL-1983 15:35	LPA01 14-JUL-1983 15:35	USER:[KNAUB]LPAWRST
VAX/VMS	KNAUB	LPAWRST 14-JUL-1983 15:35	LPA01 14-JUL-1983 15:35	USER:[KNAUB]LPAWRST

```

INTEGER PRFV(317)
DIMENSION I(1000),NEXT(1000),II(3),MIN(317),MAX(317)
LOGICAL*1 FLAG(317)
REAL*4 I,II,MINVAL,MAXVAL
REAL*8 BOUND(318),XX,PS,TPS,PTPS,PRC,PRP,C2,PBC1,PRC2,
/ PRC3,PBC4
DATA ISEED/78125/,IDIR/1/
M=0
WRITE(19,111)
111 FORMAT(1X,'LEHMANN POWER ANALYSIS FOR THE WILCOXON
/ RANK SUM TEST, LPAWRST')
WRITE(19,1)
WRITE(6,1)
1 FORMAT(1X,'ENTER NO. OF OBSERVATIONS, NO. Y')
READ(5,*)NOBS,NY
WRITE(19,*)NOBS,NY
WRITE(19,101)
7 PRINT 101
101 FORMAT(1X,'INPUT TEST STATISTIC')
READ(5,*)IX
WRITE(19,*)IX
WRITE(19,102)
PRINT 102
102 FORMAT(1X,'INPUT NO. OF REPLICATIONS')
READ(5,*)IREPS
WRITE(19,*)IREPS
100 I(1)=RAN(ISEED)
IYRNM=0
DO 105 J=1,317
FLAG(J)=.FALSE.
105 CONTINUE
MINVAL=I(1)*IDIR
MAXVAL=0
DO 10 J=2,NOBS
I(J)=RAN(ISEED)
II(1)=I(J)*IDIR
IF(II(1).GT.MAXVAL)MAXVAL=II(1)
IF(II(1).LT.MINVAL)MINVAL=II(1)
10 CONTINUE
J=J-1
NUM=J
X=NUM
NCELLS=(X/(SORT(X)))+.5
RANGE=MAXVAL-MINVAL
BOUND(1)=MINVAL
XX=RANGE/NCELLS
DO 60 J=2,NCELLS
BOUND(J)=BOUND(J-1)+XX
60 CONTINUE
BOUND(NCELLS+1)=MAXVAL
DO 50 J=1,NUM
XX=I(J)*IDIR
DO 70 JJ=1,NCELLS
IF(XX.GE.BOUND(JJ).AND.XX.LT.BOUND(JJ+1))GO TO 75
70 CONTINUE
JJ=JJ-1
75 IF(FLAG(JJ).EQ..FALSE.)THEN
MIN(JJ)=J
MAX(JJ)=J
FLAG(JJ)=.TRUE.
GO TO 50

```

```

      END IF
      II(1)=XX
      II(2)=I(MIN(JJ))*IDIR
      II(3)=I(MAX(JJ))*IDIR
      IF(II(1).LE.II(2))THEN
        NEXT(J)=MIN(JJ)
        MIN(JJ)=J
      ELSE IF(II(1).GT.II(3))THEN
        NEXT(MAX(JJ))=J
        MAX(JJ)=J
      ELSE
        PREV(JJ)=MIN(JJ)
        K=NEXT(MIN(JJ))
      20    II(1)=I(J)+IDIR
        II(2)=I(K)+IDIR
        IF(II(1).LE.II(2))GO TO 30
        PREV(JJ)=K
        K=NEXT(K)
        GO TO 20
      30    NEXT(PREV(JJ))=J
        NEXT(J)=K
      END IF
      50    CONTINUE
      L=0
      PS=1.
      IIY=0
      DO 80 JJ=1,NCELLS
        IF(FLAG(JJ).EQ..FALSE.)GO TO 80
        K=MIN(JJ)
      40    IIY=IIY+1
        IF(K.LE.NY)THEN
          L=L+1
          IYRNM=IYRNM+IIY
          PS=PS*(IIY+L-1)/100.
          IALPH='Y'
        ELSE
          IALPH='X'
        END IF
      1    WRITE(6,2)J(K),IALPH
      2    FORMAT(1X,F15.7,5X,A1)
        IF(K.EQ.MAX(JJ))GO TO 80
        K=NEXT(K)
        GO TO 40
      80    CONTINUE
      1    WRITE(6,3)IYRNM
      3    FORMAT(1X,'SUM OF Y - RANKS = ',16)
        IF(IYRNM.GE.1X)M=M+1
        IF(IYRNM.EQ.1X)THEN
          PTPS=PTPS+PS
          NP1PS=NP1PS+1
        END IF
        IF(IYRNM.LE.1X)GO TO 200
        TPS=TPS+PS
        NP1PS=NP1PS+1
        NTPS=NTPS+1
        PTPS=PTPS+PS
      200    ITRACK=ITRACK+1
        XM=M
        IF(ITRACK-ITRPS)100,201,201
      201    XHONS=NOUS
        XNY=NY

```

```

C1=IREPS
PA=XM/C1
XNTPS=JITPS
WRITE(19,*)XNTPS
XNPTPS=NPTPS
IF(TPS.EQ.0.)THEN
    ATPS=0.
ELSE
    ATPS=1PS/XNTPS
END IF
IF(PIPS.EQ.0.)THEN
    APTPS=0.
ELSE
    APTPS=PIPS/XNPTPS
END IF
C2=(XNORS+XNY)*(10.**20)
DO 203 IL=1,NY-1
    X1=IL
    C2=C2*(XNORS+XNY-X1)/100.
203 CONTINUE
    WRITE(6,*)C2
    XNYF=1.
    PB=A1PS*PA
    PBP=APTPS*PA
    PBC1=(2.**30)*(10.**20)
    PBC2=(2.**30)*PBP
    PBC3=(10.**2)*(2.**(NY-60))
    PBC4=PBC3/C2
    POWER=PBC1*PBC2*PBC4
    WRITE(19,*)IY,L,PB,C2
    IF(ATPS.EQ.0.)THEN
        Pd=0.
    ELSE
        PB=POWER*(PB/PBP)*(XNTPS/XM)
    END IF
900 Pd=1.0-PB
    WRITE(6,*)PA,POWER,PB
    WRITE(19,*)PA,POWER,PB
    STOP
END

```

APPENDIX B

ACKNOWLEDGEMENTS

Thanks to Lynne Grile and Gerard Petet for developing the programming necessary to generate graphs 1 and 2. Keith Haycock (White Sands) constructed graphs 3-7, and Dr. Larry Armijo (White Sands - KENTRON) wrote the computer program to provide analytical solutions used in graphs 3-7. Also, thanks to Jeffrey Greenhill for providing a customized sorting routine for the author's simulation. Carlyle Comer and Frank Lawrence were helpful in obtaining massive quantities of needed CPU time on the USALOGC's VAX-11/780. Finally, thanks to other analysts for helpful conversations and/or influence.

APPENDIX C

REFERENCES

1. Conover, W. J., Practical Nonparametric Statistics, 2 ed, John Wiley & Sons, 1980.
2. Downs, R. S., P. C. Cox, "The Probability of Motor Case Rupture," ARO Report 75-2.
3. Juritz, J. M., J. W. F. Juritz and M. A. Stephens, "On the Accuracy of Simulated Percentage Points," Journal of the American Statistical Association, 78 (June 1983).
4. Knaub, J. R., Jr., "Design of a Multiple Sample Westenberg Type Test for Small Sample Sizes," ARO Report 82-2.
5. Knaub, J. R., Jr., Appendix D, US Army Manpower Nonavailability and Indirect (Unit Related) Productive Factors - Final Report, US Army Logistics Center, August 1983 - not released
6. Knaub, J. R., Jr., L. M. Grile and G. Petet, "Analyzing n Samples of 2 Observations Each," ARO Report 83-2.
7. Lehmann, E. L., "The Power of Rank Tests," Annals of Mathematical Statistics, 24 (1953), 23-43.

ADDENDUM

Multiple applications of this test can be used to compare two levels of a factor under a number of conditions. If, for example, manufacturer A produces a machine which is suspected to have higher reliability under most scenarios than a similar machine made by manufacturer B, then under each of the γ scenarios, m_i is the sample size of A's machines and n_i is the sample size of B's machines, for $i = 1$ to γ . PA_i and PB_i can be calculated for each of the scenarios. Consider $0 \leq a \leq \gamma$ and $0 \leq b \leq \gamma$:

PA is the probability of a or more PA_i 's being less than p_A
 $(i = 1; \gamma)$, when H_0 is true.

PB is the probability of b or more PB_i 's being less than p_B
 $(i = 1; \gamma)$, when H_1 is true.

Therefore,

$$\begin{aligned} \text{and } PA &= \sum_{x=a}^{\gamma} \binom{\gamma}{x} P_A^x (1 - P_A)^{\gamma-x} \\ PB &= \sum_{x=b}^{\gamma} \binom{\gamma}{x} P_B^x (1 - P_B)^{\gamma-x} \end{aligned}$$

P_A and P_B are chosen to be reasonable considering sample sizes for each of the γ cases.

If $\frac{PA}{PB} = 1$ then the evidence shows that, in general, the true state of nature is just as likely to be equivalent to H_1 as H_0 .

If $\frac{PA}{PB} = 2$ then the evidence indicates that, in general, the true state of nature is twice as likely to be equivalent to H_0 as H_1 . If PA and PB are small, then the indication is only that the true state of nature is closer to H_0 than H_1 , although possibly not very close to either.

(Note that another paper in this conference, "Numerical Validation of Tukey's Criteria for Clinical Trials and Sequential Testing," by C. R. Leake, also deals with this type of problem, and was of interest to this author.)

At this time, this methodology is being used to determine whether survey data from a presumably less reliable source is compatible with a presumably superior data source. Difficult to obtain data on U.S. Army warehousing activities have, as one obvious characteristic, a very flat "peak." Therefore, a sample median value can be changed drastically by the addition or deletion of one data point. If the secondary data source proves to provide values distributed closely enough to that of the primary source, the advantage of including this source may outweigh the disadvantage. The current situation is more complex than this. However, some results employing the methodology of this addendum have been realized.

ADDENDUM 2

Two approximations for the power of this test which apparently are good for a wide range of normal alternative hypotheses are to be found in E. L. Lehmann, Nonparametrics: Statistical Methods Based on Ranks, Holden-Day, 1975. Although restricted to normal alternatives in the format in which they are written, these approximations can be used to extend the tables given here to larger n and m . The easier of the two approximations to apply, in its simplest form, is found on page 73 of the above reference and is essentially as follows:

$$\text{power} \approx \Phi \left[\sqrt{\frac{3mn}{(m+n+1)\pi}} \frac{\mu_A - \mu_B}{\sigma} \chi_{1-\alpha} \right]$$

where in our case we have $(\mu_A - \mu_B)/\sigma \approx 0.610$.

Note that in the example in the main body of this paper ($m = 19, n = 11$), that this approximation gives power ≈ 0.60 , which is consistent with what was shown earlier.

COMPLEX DEMODULATION - A TECHNIQUE FOR ASSESSING
PERIODIC COMPONENTS IN SEQUENTIALLY SAMPLED DATA

Helen C. Sing, Sander G. Genser, Harvey Babkoff*,
David R. Thorne, and Frederick W. Hegge
Department of Military Medical Psychophysiology
Division of Neuropsychiatry
Walter Reed Army Institute of Research
Washington, D. C. 20307
*Bar-Ilan University
Ramat-Gan, Israel

ABSTRACT. Circadian and other rhythmic components in data obtained from a sleep deprivation study are detected and characterized by complex demodulation (CD). The output of this analytical technique yields both frequency and time domain representation of each periodic component of interest. Non-stationarity introduced by an experimental treatment such as progressive sleep loss, may be observed and quantified.

The analytical results provide a common basis of comparison for data as diverse as cognition responses from a performance assessment battery (PAB), moodscale scores, and physiological data such as oral temperature.

The procedure operates on the entire data set and variance accounted for by each component may be calculated.

I. INTRODUCTION. Our laboratory has been involved in probing the problems dealing with sleep discipline that are directly pertinent to soldiers in battlefield situations. In the process of conducting a series of experiments of continuous sleep deprivation over 48 and 72 hours, a massive amount of data has been collected [1]. These data sets are of such diverse nature as electrocardiography, actigraphy based on measurement of movement on a non-dominant wrist, oral temperature, self scored reports of mood/activation and cognitive/visual difficulties, a computerized battery of performance assessment tasks, and a computerized lexical decision task.

Taken in synchrony, these data have in common the characteristic of equal interval time sampling, whether imposed or extractable, that is to say, temperature, test results, self reports are taken at scheduled intervals while continuously recorded data such as electrocardiographs and actigraphs may be extracted with the same time intervals.

How can their commonality in time be exploited so that the subtle changes from an intervention, i.e., sleep deprivation, may be observed in each type of data, and what are the relationships among data sets.

Standard statistical analyses such as ANOVA, MANOVA, etc., are helpful in pointing out general significance or non significance among data sets but are not helpful in pinpointing exact locations of similarities or differences in

the time oriented dimension. Other time series analyses such as Fourier transforms, auto or cross correlations are global in nature and again do not yield local parameters.

Towards this end, we have taken a technique more commonly used in signal analysis and adapted it to our specific needs so that the resulting analysis provides information on an epoch by epoch basis over the entire sampling period [2,3]. We have made some simplifications, perhaps taken some liberties, but our emphasis has been more on practical applications rather than on mathematical rigor. Nevertheless, our analyses have yielded faithful approximations to the original data sets and have provided us with the local parameters of power, amplitude, phase, and remodulate for each epoch [4].

II. THE METHOD OF COMPLEX DEMODULATION (CD).

The data set comprising measurements taken in equal time increments (epochs), is given as:

$$X(t) = x_1, x_2, \dots, x_N \quad (1)$$

where each x element is the value of the measurement at that epoch and x_1 is the first epoch value at time, $t = 0$. The epoch length may be 1 minute, 15 minutes, 1 hour, etc. Although oral temperature was taken hourly and ECG and actigraph epoch lengths were less than 1 min for the data collected here, the computerized tasks were given at alternate hour intervals. The epoch length used in the CD analysis for all data except the computerized tasks was 1 hr, while a 2 hr epoch was used for the computerized tasks. However, comparisons of all data types were standardized to 2 hr epochs.

Subtraction of the data series' mean value from each epoch datum yields the set:

$$Y(t) = y_1, y_2, \dots, y_N \quad (2)$$

where

$$y_i = x_i - \frac{1}{N} \sum_{n=1}^N x_n \quad (3)$$

$$i = 1, 2, \dots, N$$

The new data set oscillates around the mean level or what is commonly referred to as the "zero frequency".

Time series are implicitly infinite in length, but actual analysis of data requires a finite set of data and hence we are faced with abrupt truncation at the beginning and end of the data set which has consequences of "end effects" resulting in distortion of local parameters at these locations

after analysis. Our experience indicates that these end effects may be minimized and or eliminated by extending the data sets at both these locations in the following way:

$$Z(t) = y_m, y_{m-1}, \dots y_1, y_1, y_2, \dots y_N, y_N, y_{N-1}, \dots y_{N-k} \quad (4)$$

where

$$\begin{aligned} m &= \text{number of folded-out data epochs} \\ k &= m - 1 \end{aligned}$$

This is reasonable in light of other alternatives, one of which adds zeroes to both ends [5]. The number of data points folded-out varies according to the length of the data set. Our rule has been to use 20% of the total values if the series is long (>100 epochs), and 5%, if a shorter segment.

All subsequent mathematical operations are made on the folded-out series. However, the final output retains only the parameters of the original epochs for statistical analysis and display.

Mapping of each data value to the complex domain follows with generation of real (re) and imaginary (im) components for each epoch in which the arguments of the respective functions contain the frequency to be elicited. These functions are:

$$z_i(\text{re}) = y_i \cdot \cos 2\pi f_j t/s \quad (5)$$

$$z_i(\text{im}) = y_i \cdot \sin 2\pi f_j t/s \quad (6)$$

where

$$i = 1, 2, \dots, N + 2m \quad (\text{indexed for extended data set})$$

$$\begin{aligned} f_j &= \text{jth frequency selected for demodulation,} \\ j &= 1, 2, \dots, s/2 \end{aligned}$$

$$t = i - 1$$

$$s = \text{number of epochs sampled in the chosen period, } T$$

For example, if period, $T = 24$ hr, frequency to be demodulated = 3 cycles, and sampling rate = 2/hr, then

$$f_j/s = 3/(24 \cdot 2) = 3/48$$

Since our procedure involves incremental sampling time of equal intervals, then t increments by 1 from time zero, which corresponds to the first data point.

Implicit in each datum is the collective value of all inherent frequencies contained in the series. Multiplication by the sine and cosine terms preserves not only the frequency demodulated, but also generates additional frequencies from the sums and differences of products between the modulating frequency and the inherent frequencies. This transforms the data from the time to frequency domain and visualized as a Fourier spectrum, places the frequency being demodulated at the zero frequency position. In a step-wise process, each frequency of the period chosen (in our cases, we have generally used the circadian period which is the 24 hr cycling most common to man) may be individually demodulated.

For a given data set, the highest frequency demodulable is the Nyquist frequency [6] and is equal to one-half the sampling rate, i.e., sampling at hourly intervals in a 24 hr period allows demodulation of frequencies 1 through 12 per day. This limitation is due to discrete equal sampling intervals in which frequencies higher than the Nyquist are enveloped by lower frequencies with which they coincide at crossover points, and are therefore "aliased" and not true frequencies.

Extraction of the desired frequency at the zero frequency position necessitates exclusion of not only the sums and differences of the products mentioned above, but also of other noise constituents. This is accomplished by a filter which is moved sequentially along both sine and cosine components of the series first in a forward pass, then a reverse pass and the entire process repeated. The forward pass causes a shift of one in the data set which is corrected by the reverse pass, thereby preserving true phase values. The filter employed in this process is exponential and consists of two parts:

$$F_1 = (A^2 + B^2)^{1/2} \quad \text{Part 1} \quad (7)$$

$$F_2 = e^{-\alpha} \quad \text{Part 2} \quad (8)$$

where

$$A = 1.0 - e^{-\alpha} \cos 2\pi\gamma/s \quad (9)$$

$$B = e^{-\alpha} \sin 2\pi\gamma/s \quad (10)$$

and γ = gain factor (variable from 0.1 to 0.9)

$$\alpha = 2\pi/s$$

s = number of epochs sampled in the chosen period

The gain factor, γ , may be varied from 0.1 to 0.9 depending on the magnitude of the original data values i.e., smaller values require higher gain. Direct comparisons of different records require that the same gain factor be used.

A new value is obtained for each datum in the series as a consequence of filtering so that for the forward pass:

$$z'_0 = z_0 = 0$$

$$z'_i(\text{re}) = z_i(\text{re}) \cdot F_1 + z'_{i-1}(\text{re}) \cdot F_2 \quad (11)$$

$$z'_i(\text{im}) = z_i(\text{im}) \cdot F_1 + z'_{i-1}(\text{im}) \cdot F_2 \quad (12)$$

where $i = 1, 2, \dots, N + 2m$ (indexed for extended data set)

and for the reverse pass:

$$z''_{N+2m} = z'_{N+2m}$$

$$z''_{i-1}(\text{re}) = z'_{i-1}(\text{re}) \cdot F_1 + z''_i(\text{re}) \cdot F_2 \quad (13)$$

$$z''_{i-1}(\text{im}) = z'_{i-1}(\text{im}) \cdot F_1 + z''_i(\text{im}) \cdot F_2 \quad (14)$$

where $i = N + 2m, N + 2m - 1, \dots, 2$

The low pass characteristics of this filter allow passage of power at and near the zero frequency in the spectrum while excluding other frequencies. Inevitably, there will be some "leakage" of power from frequencies located adjacent or near to the zero frequency position. For this reason, in our analysis of human data where the strongest frequency is the circadian (1 cycle per 24 hr), epoch values obtained from remodulates (to be defined shortly) of frequency 1, are subtracted from their corresponding values in the folded-out data set before demodulating in the usual way for all subsequent frequencies.

In practice, the filter operation involves summing a proportion of each epoch value with a proportion of the previous one. The outputs of each filter pass are used as new inputs for the next pass in the reverse direction.

The final outputs from the filter operations are used for computing the local parameters or properties of each epoch. These are:

$$\text{Power: } P_1 = 2.0[z''_1{}^2(\text{re}) + z''_1{}^2(\text{im})] \quad (15)$$

$$\text{Amplitude: } P_1^{1/2} \quad (16)$$

$$\text{Phase: } \phi_1 = \arctan[z''_1(\text{im})/z''_1(\text{re})] \quad (17)$$

$$\text{Remodulate: } R_1 = 2.0[z''_1(\text{im}) \sin 2\pi f_j t/s + z''_1(\text{re}) \cos 2\pi f_j t/s] \quad (18)$$

where $i = 1, 2, \dots, N$ (indexed for original data set)

f_j = demodulated frequency, j

$j = 1, 2, \dots, s/2$

$t = i - 1$

s = number of epochs sampled in the chosen period

The remodulate values, after truncation of the folded-out epochs at the beginning and terminus of the data series, comprise a smoothed function of the desired demodulated frequency with the proper phases. The remodulates are used in all subsequent comparisons. Peak and trough amplitudes and their corresponding real times may be determined over the entire length of the series for every frequency demodulated. Actual length (in hours) of the circadian period may be calculated either from peak to peak, trough to trough, or zero cross over points depending on the interest. Instantaneous changes in phase (from phase plots) signal changes in period length, i.e., frequency, and may be detected from records taken over several cycles.

III. ILLUSTRATIONS.

Graphical representations best illustrate the method and results of the various types of data we have analyzed.

Figure 1 (top) depicts the original data series and (bottom) illustrates how the data set is folded out at the beginning and terminal ends.

Figure 2 is the representation of transformation to the complex domain of sine and cosine components of the data set with the circadian (1 cycle/24 hr) filtered output from these superimposed in heavy outline.

Figure 3 summarizes the CD procedure, as plots, from the original input data to output parameters of amplitude, phase, and remodulate of the circadian component along the time scale in epoch intervals.

In our 72 hr sleep deprivation study, the subjects' oral temperatures were converted to z-scores to facilitate comparisons across subjects. Since a strong linear component with negative slope was observed over the 3 days' running, CD was performed on the residual (fig. 4) from the least squares regression of the z-scores. Frequencies of 1 through 12 cycles per day (cpd) were demodulated and plots of their remodulates generated. Some of these plots are presented here. Figure 5 shows the circadian with its daily rhythmic cycling of temperature rising slowly during the morning, peaking in early evening and then dropping to its lowest point usually between 2 and 4 A. M. Moreover, there is broadening of wave shape on the 2nd and 3rd days of sleep deprivation, indicating changes in phase and period. There is an accelerated decline at the close of Day 2 in the raw data and this is reflected in the steeper trough for the circadian rhythm. The remodulate of 2 cpd shown in Figure 6 may represent the post-prandial dip that is sometimes seen as bimodal in the raw data. Figures 7 and 8 are the 4 cpd and 12 cpd components respectively. Increase in amplitudes of higher frequencies components may be signals of intrinsic system instability i.e., subjects' reports of feeling cold despite normal room temperature, of appetite loss, and of eating and drinking. Summation of the circadian remodulate with the 2 cpd is presented in Figure 9 and of all remodulates in Figure 10. Note that in the final summation (Fig. 10), there are no 'end effects' distortion and the summed remodulates follow the raw data shaping almost identically.

The variance accounted for by each individual frequency demodulated and by cumulative frequencies are given as R-squared values in Table 1. These R^2 values are derived from regression of each remodulate with the original detrended data set. The total power from the summed data epochs for each frequency is listed in Table 2 along with the cross correlates of each remodulate frequency with the detrended data set. The cross correlates are measures of peaks and troughs correspondence between the original detrended series and each of the remodulate frequency.

Other applications of complex demodulation have been to "throughput" measures of performance [7], during the same sleep-deprivation studies. This is a single-valued performance index derived from the ratio of accuracy to mean reaction time and describes the rate at which the subject gives "effective" performance as a function of time on task. There is increasing performance deficit over time as seen in plots of the original data (Figure 11), however, the rhythmic components are evident and are elicited from the CD procedure (Figure 12). A comparison of the circadian remodulate of the PAB scores with the subject's oral temperature is shown in Figure 13. Note the phase difference between performance and temperature with the latter leading performance. CD of scores from a mood scale check list taken by the subjects before each administration of PAB, indicates the same decline in activation and affect over the time the subject is sleep deprived. This is shown in Figure 14. However, the capacity to maintain the circadian rhythm is still apparent as is seen in Figure 15.

Finally, scores from a five point self-scoring computerized questionnaire containing fifty six queries relating to hallucinations, delusions, and illusions [8], grouped as to either: 1) cognitive (C), 2) Visual perceptual (V), 3) non-visual perceptual (N), are analyzed by CD and the results for the circadian rhythm are shown as remodulates along with oral temperature in Figure 16. Note that at the beginning of the study, circadian rhythmicity for non-visual perceptive problems and cognitive difficulty is not well defined since the subject's response was mostly at the same low level to those factors over the first 30 hours or so. On the other hand, visual perceptual problems are rhythmic, but out of phase with temperature, which is logically reasonable, that is to say, when the subject is at the peak of his cycle and feeling generally well or better, he experiences no visual problems of perception. Note that the other measures of cognitive difficulty and non-visual perceptual problems when finally reported as occurring also vary rhythmically but again out of phase with oral temperature.

IV. CONCLUSIONS.

The entire procedure of CD is computerized. There are other refinements such as use of a spline fitting program [9] to calculate for missing values and also to obtain finer resolution of times of peak or trough occurrences by interpolation between epochs. We have in addition, set strict criteria for accepting frequencies demodulable within the Nyquist frequency range as "true" or noise elements by eliminating those frequencies whose peak amplitudes are not within the ten percent population of highest peak values.

TABLE 1

VARIANCE CONTRIBUTED BY EACH FREQUENCY DEMODULATED

<u>Frequency(cpd)</u>	<u>R²</u>
1 (circadian)	0.5804
2	0.0336
3	0.1531
4	0.1362
5	0.1134
6	0.0934
7	0.0707
8	0.0678
9	0.0502
10	0.0540
11	0.0591
12	0.0442

<u>Cumulative Frequencies</u>	<u>R²</u>
(Thru)	
2	0.7003
3	0.7618
4	0.7729
5	0.8002
6	0.8123
7	0.8329
8	0.8473
9	0.8621
10	0.8751
11	0.9002
12	0.8625*

* Addition of the 12 cpd component decreased total variance accounted for.

TABLE 2

POWER OF EACH FREQUENCY AND CROSS CORRELATION OF REMODULATES
WITH ORIGINAL DETRENDED DATA

<u>Frequency (cpd)</u>	<u>Power</u>	<u>Cross Correlate of Remodulate With Detrended Data</u>
1	47.5827	36.8280
2	5.1613	3.1835
3	3.7997	5.9312
4	5.0920	6.5174
5	4.3713	4.7435
6	2.3933	3.9863
7	1.8397	3.3279
8	2.3014	2.6159
9	4.0929	2.7306
10	1.3038	1.7613
11	2.5519	2.7066
12	6.4367	5.2931

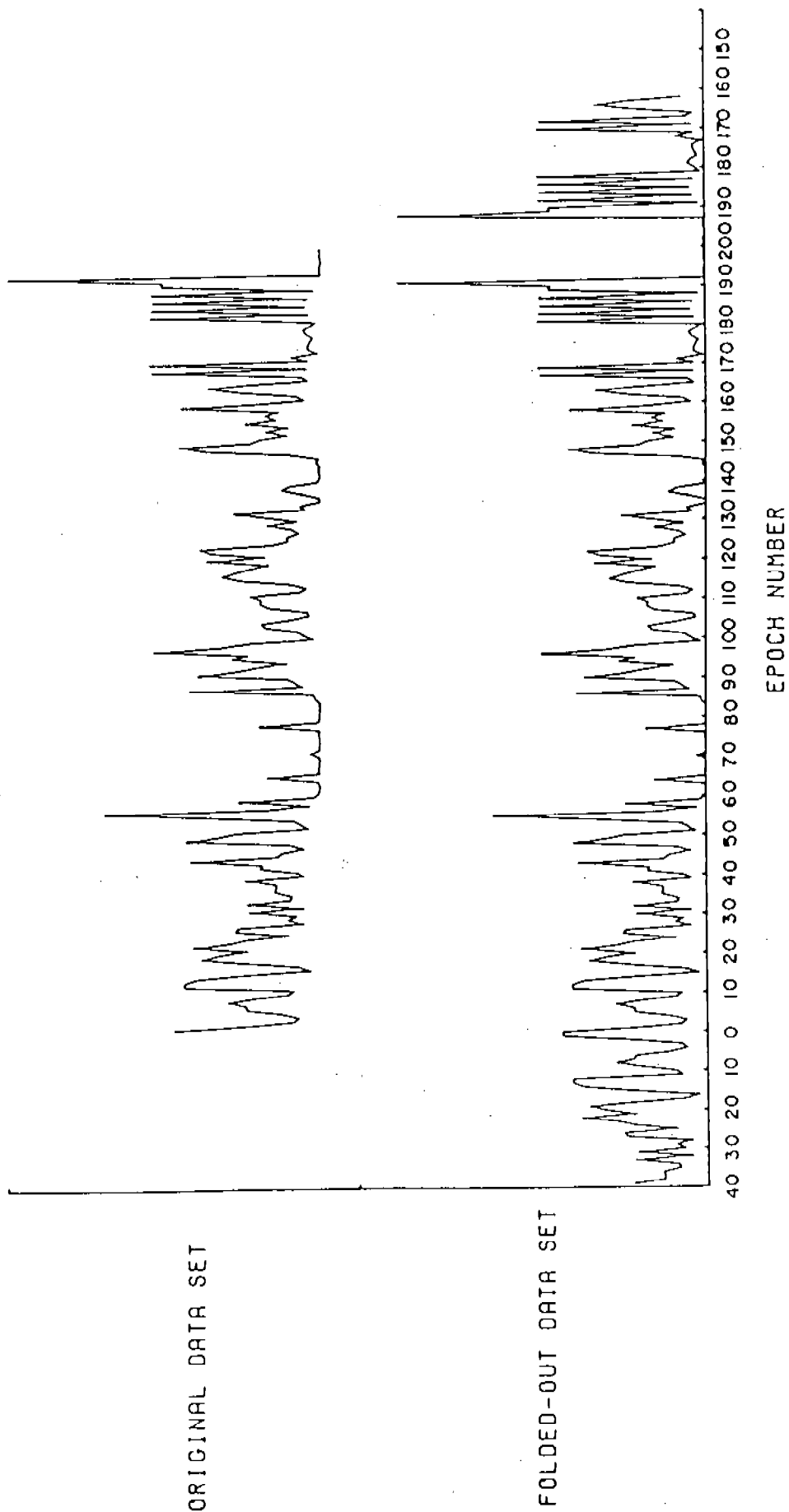


FIGURE 1

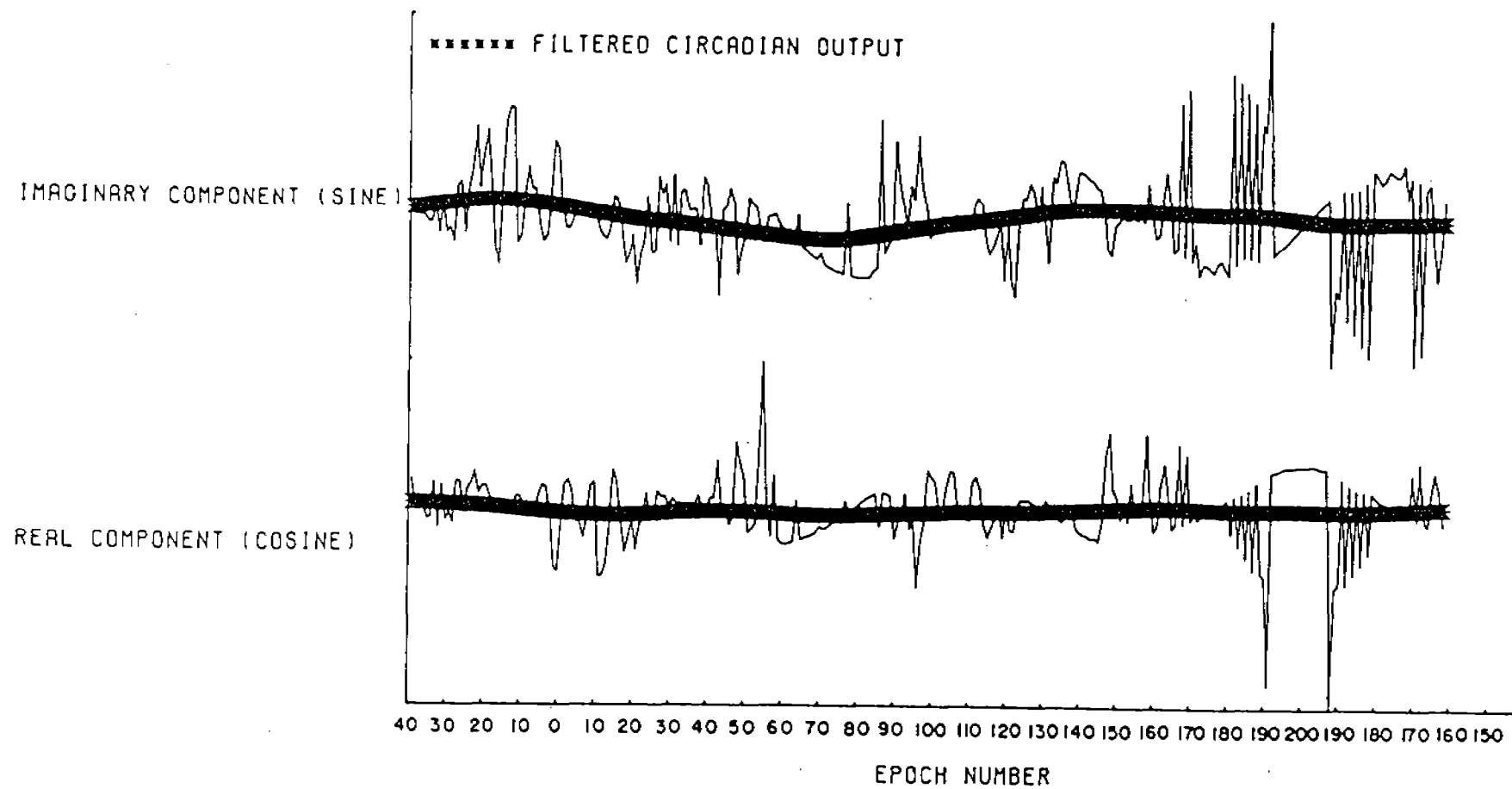


FIGURE 2

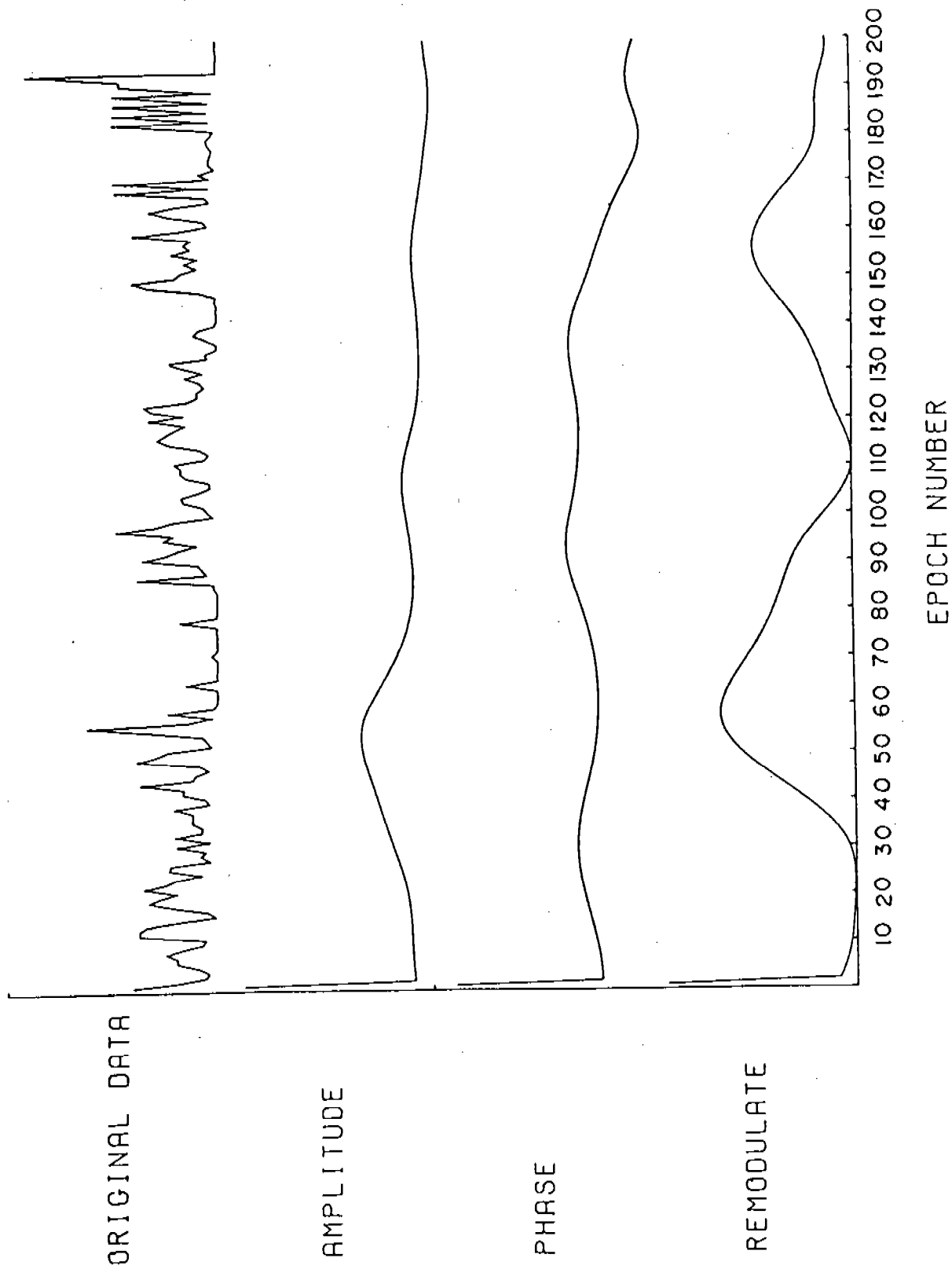


FIGURE 3

RAW DATA OF RESIDUAL (MINUS LINEAR) OF ZTEMP

PLOT OF Y* $\bar{X}$ SYMBOL USED IS O

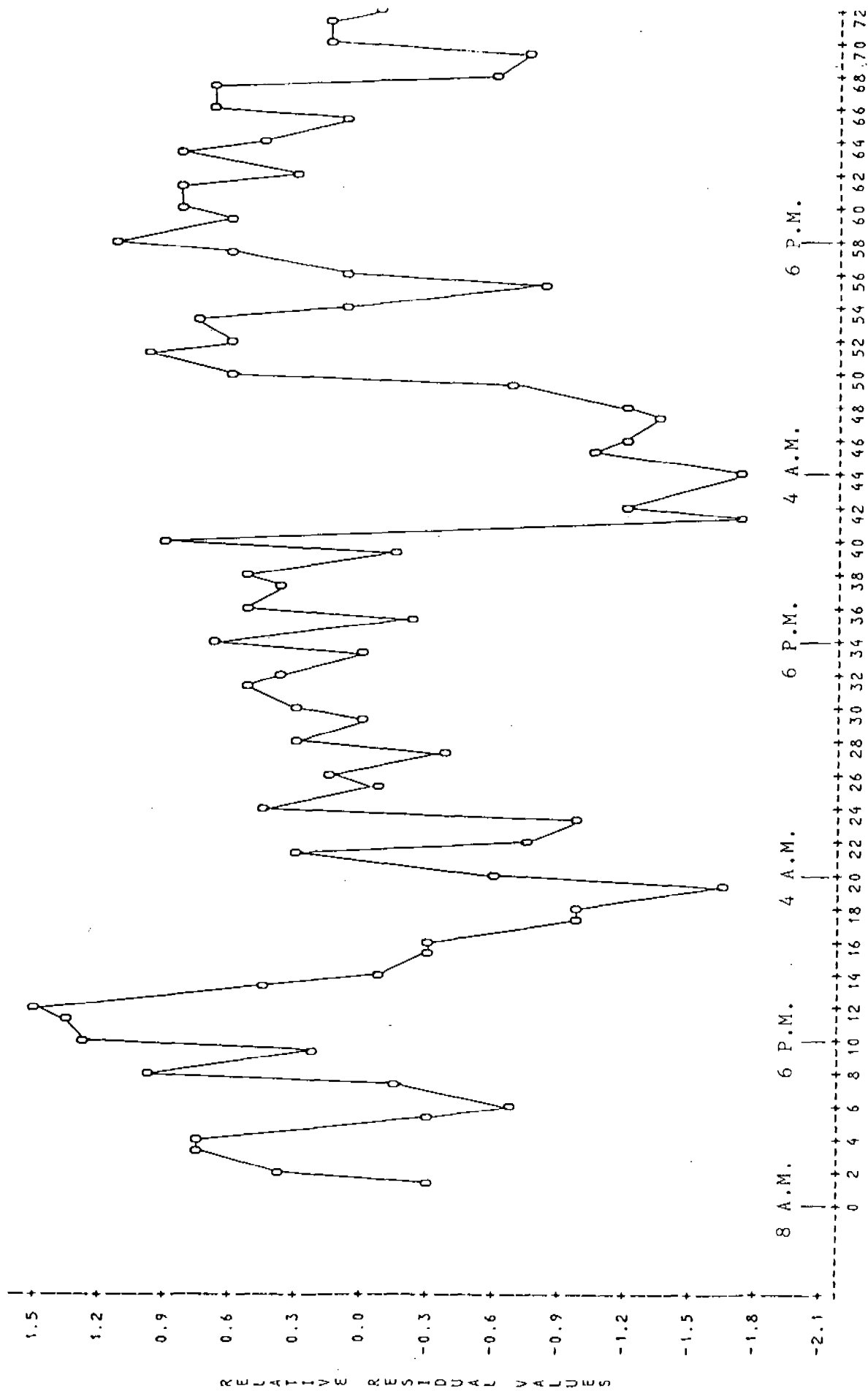
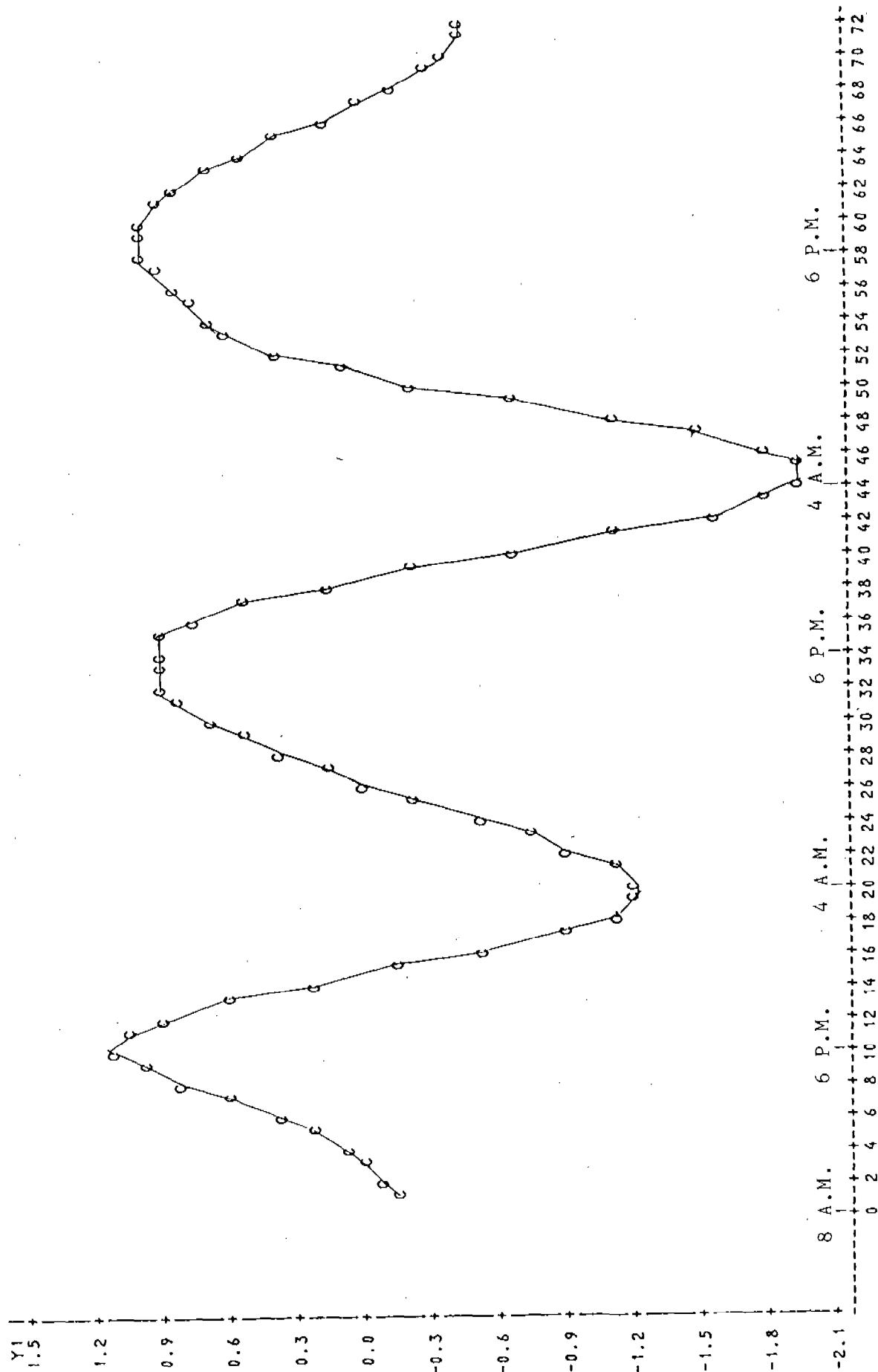


FIGURE 4

ZTEMP

CIRCADIAN REMODULATE OF RESIDUAL -

PLOT OF Y1XX SYMBOL USED IS C



SEQUENTIAL 1 HR INTERVALS

FIGURE 5

TWELVE HR COMPONENT (2CPD) OF RESIDUAL -

ZTEMP

PLOT OF Y2XX SYMBOL USED IS 2

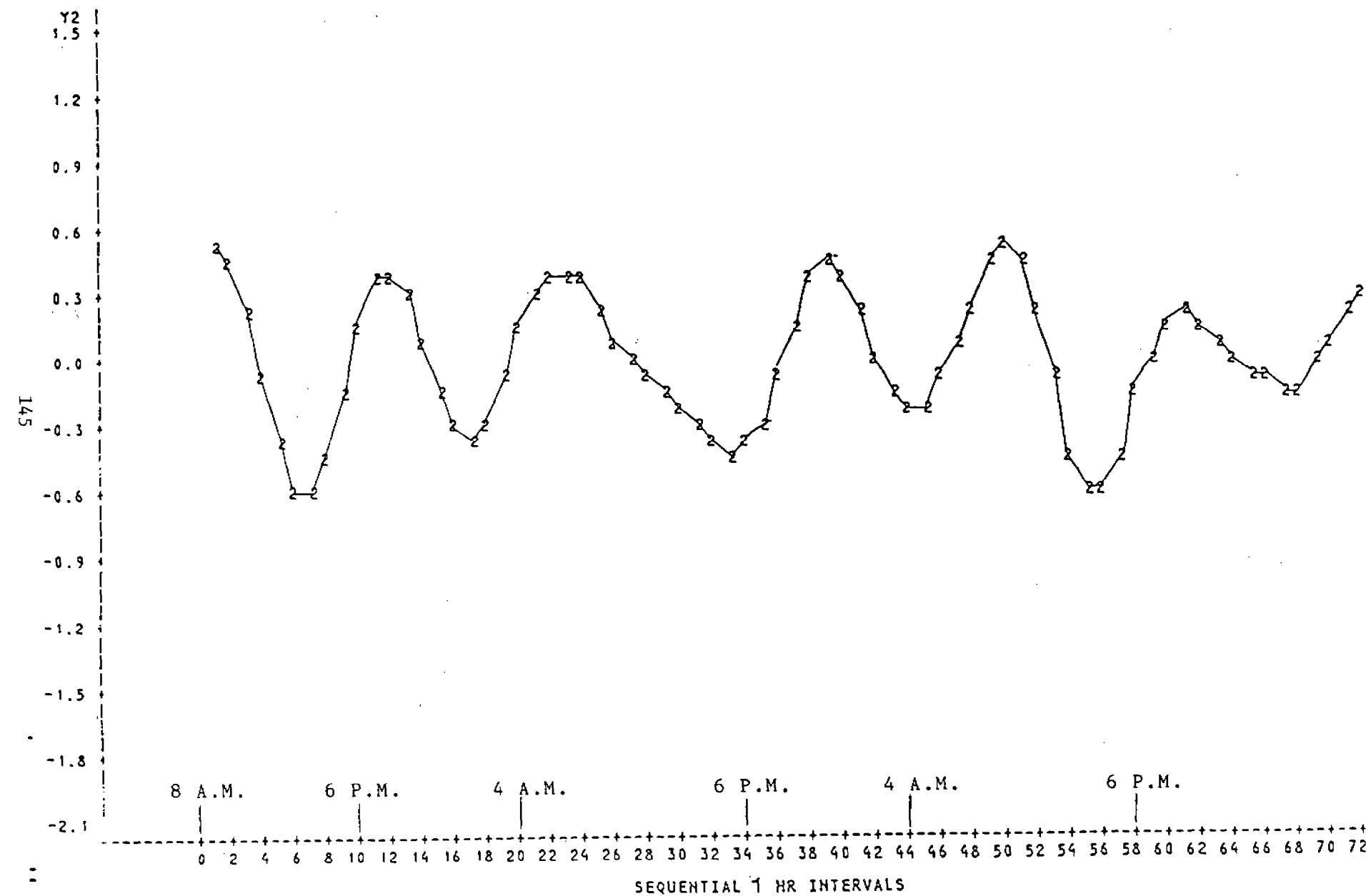
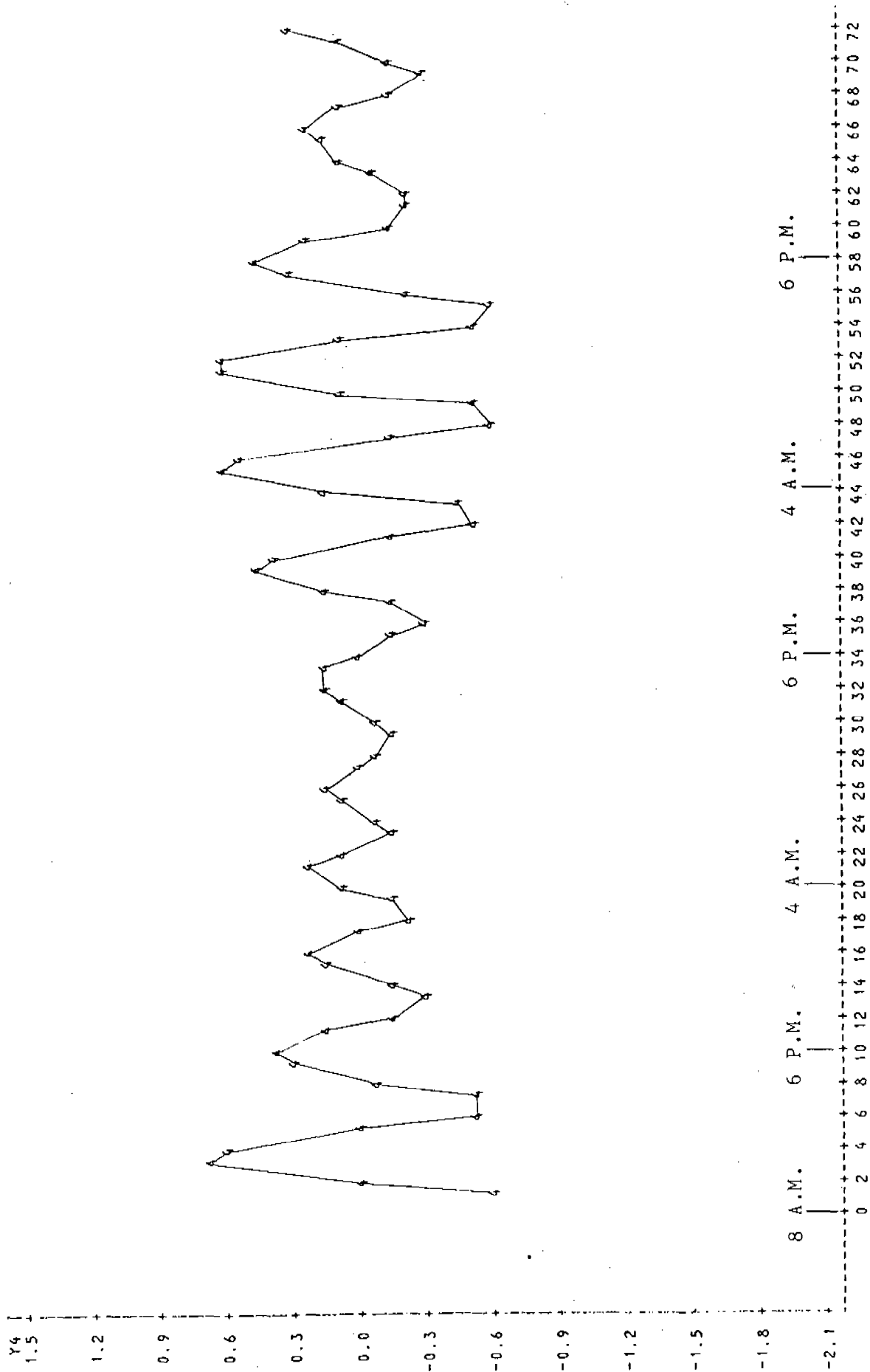


FIGURE 6

ZTEMP

SIX HR COMPONENT (4CPD) OF RESIDUAL -

PLOT OF Y4XX SYMBOL USED IS 4



SEQUENTIAL 1 HR INTERVALS

FIGURE 7

ZTEMP

TWO HR COMPONENT (12CPD) OF RESIDUAL -

PLOT OF Y12XX SYMBOL USED IS 2

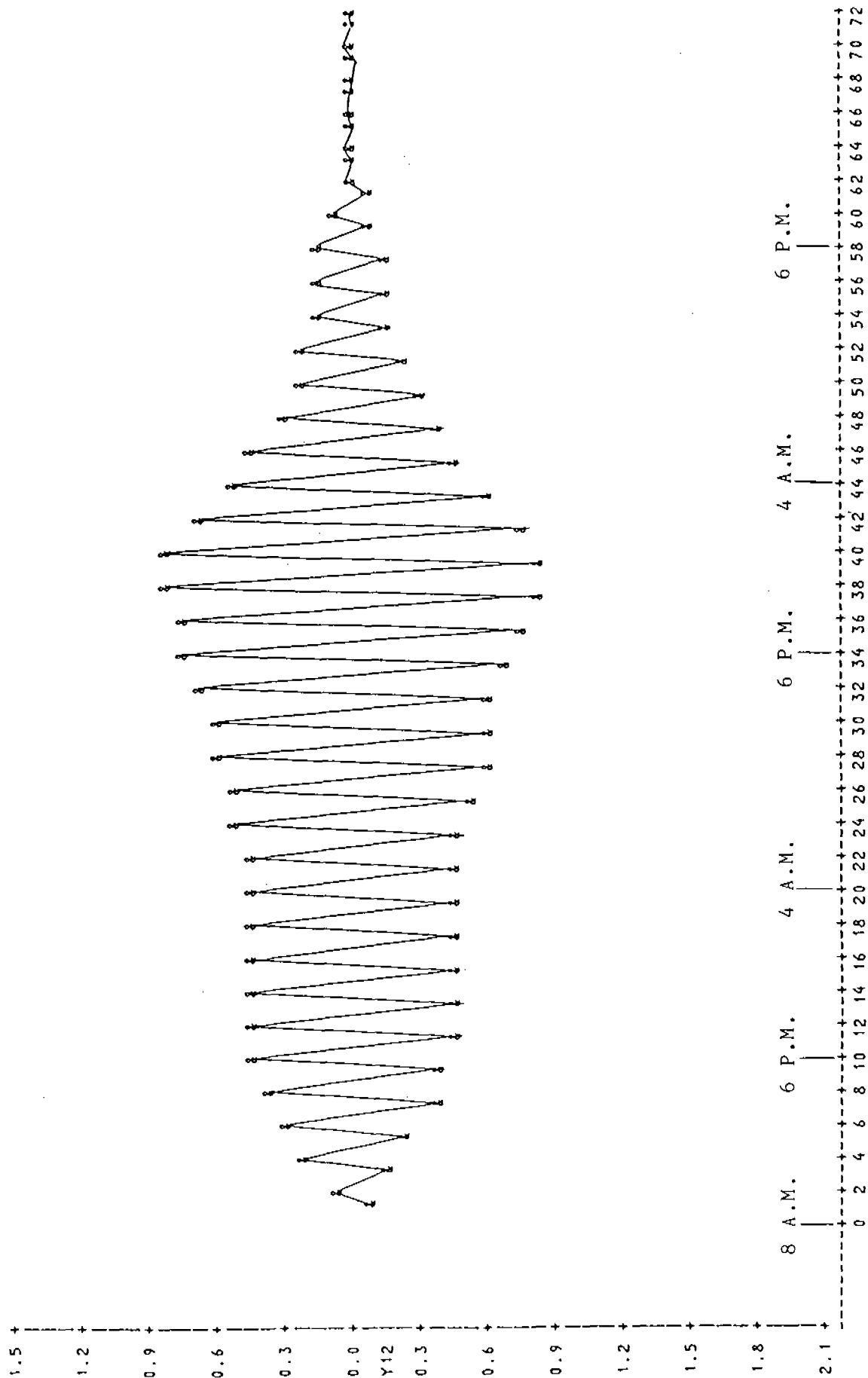


FIGURE 8

SUM OF CIRCADIAN & ZCPD REMODULATE -
PLOT OF YSUM12*X SYMBOL USED IS +

ZTEMP

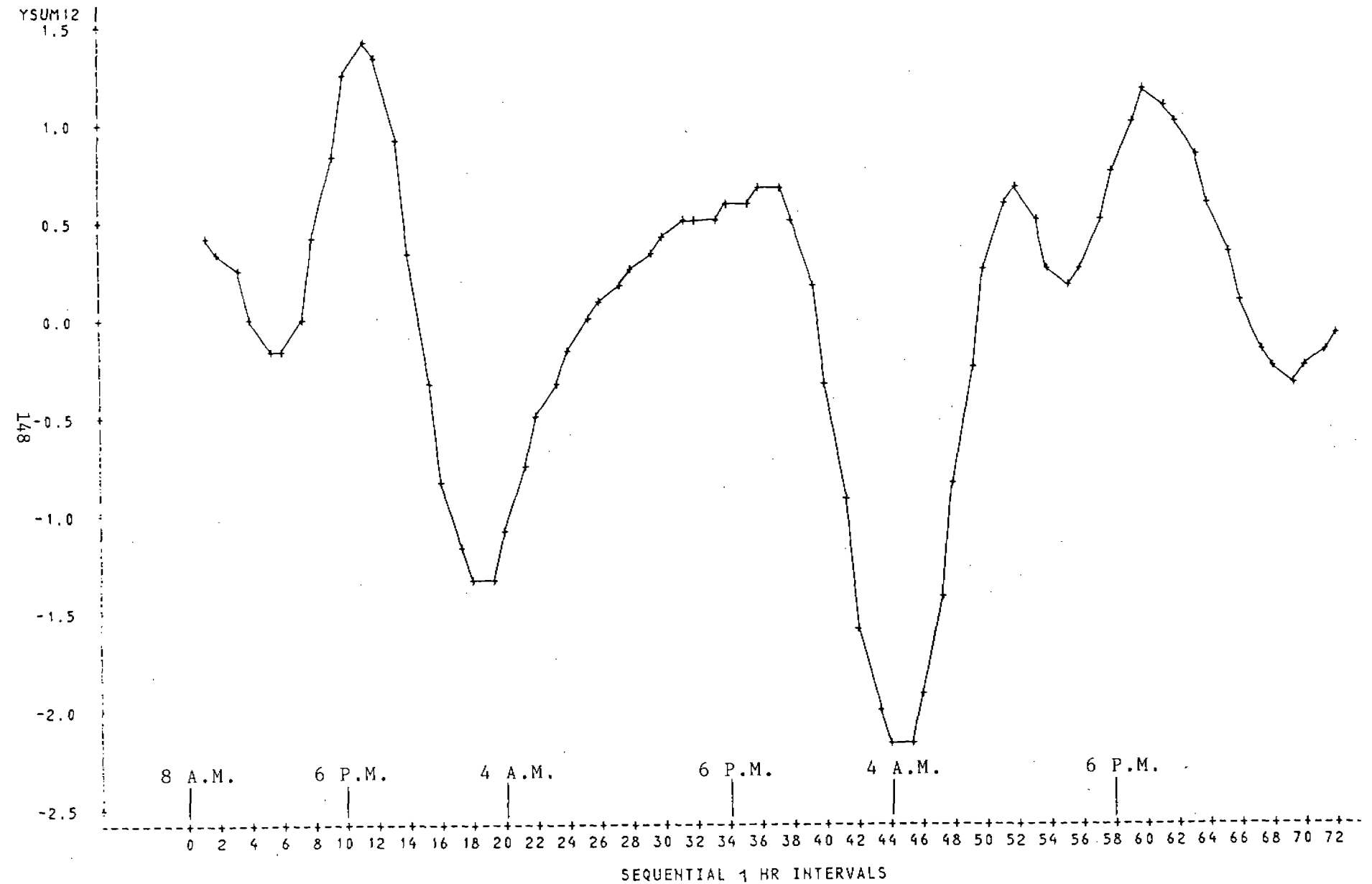
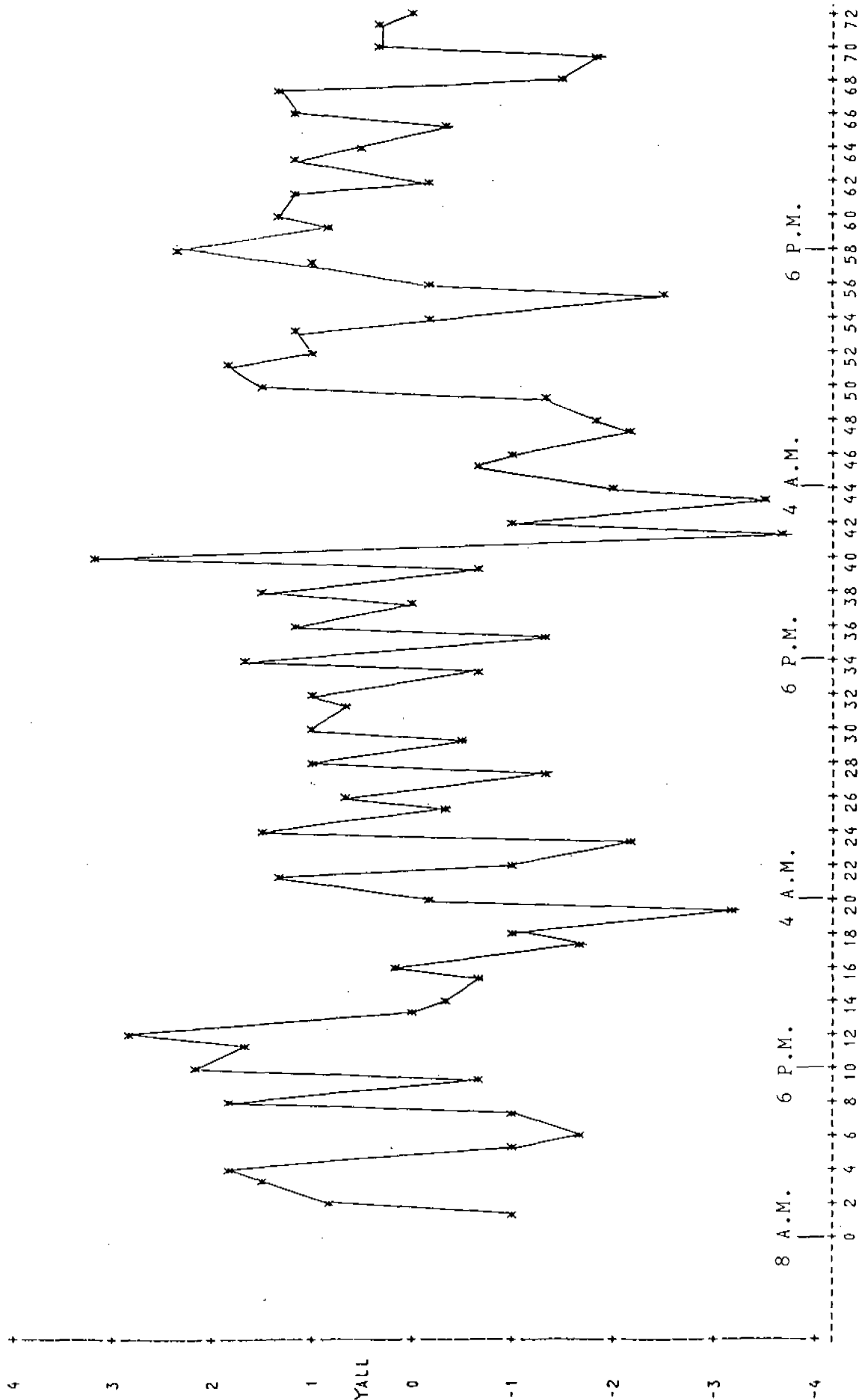


FIGURE 9

SUM OF CIRCADIAN THRU 12CPD REMODULATE -

ZTEMP

PLOT OF YALL*
SYMBOL USED IS *



SEQUENTIAL 1 HR INTERVALS

FIGURE 10

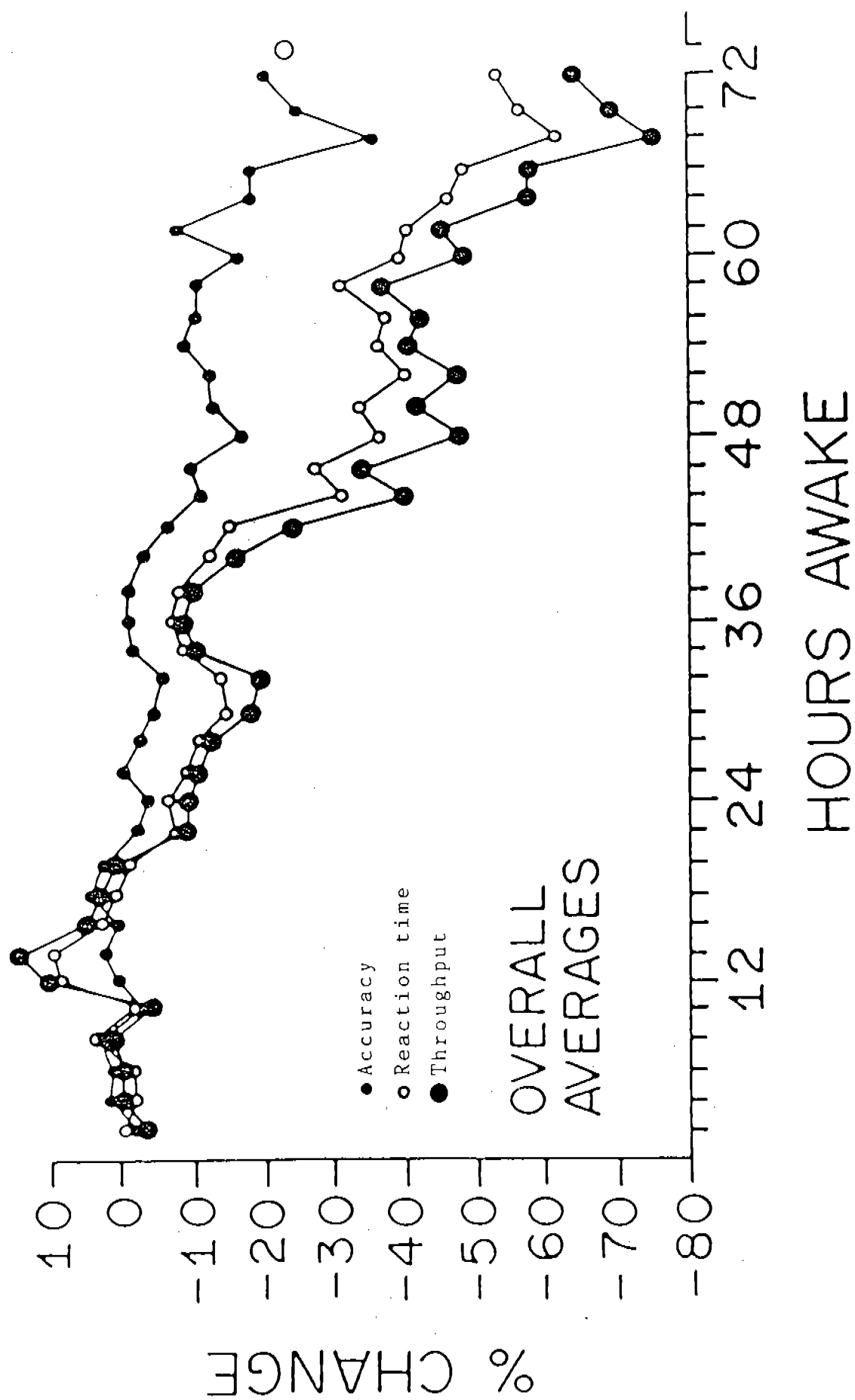


FIGURE 11

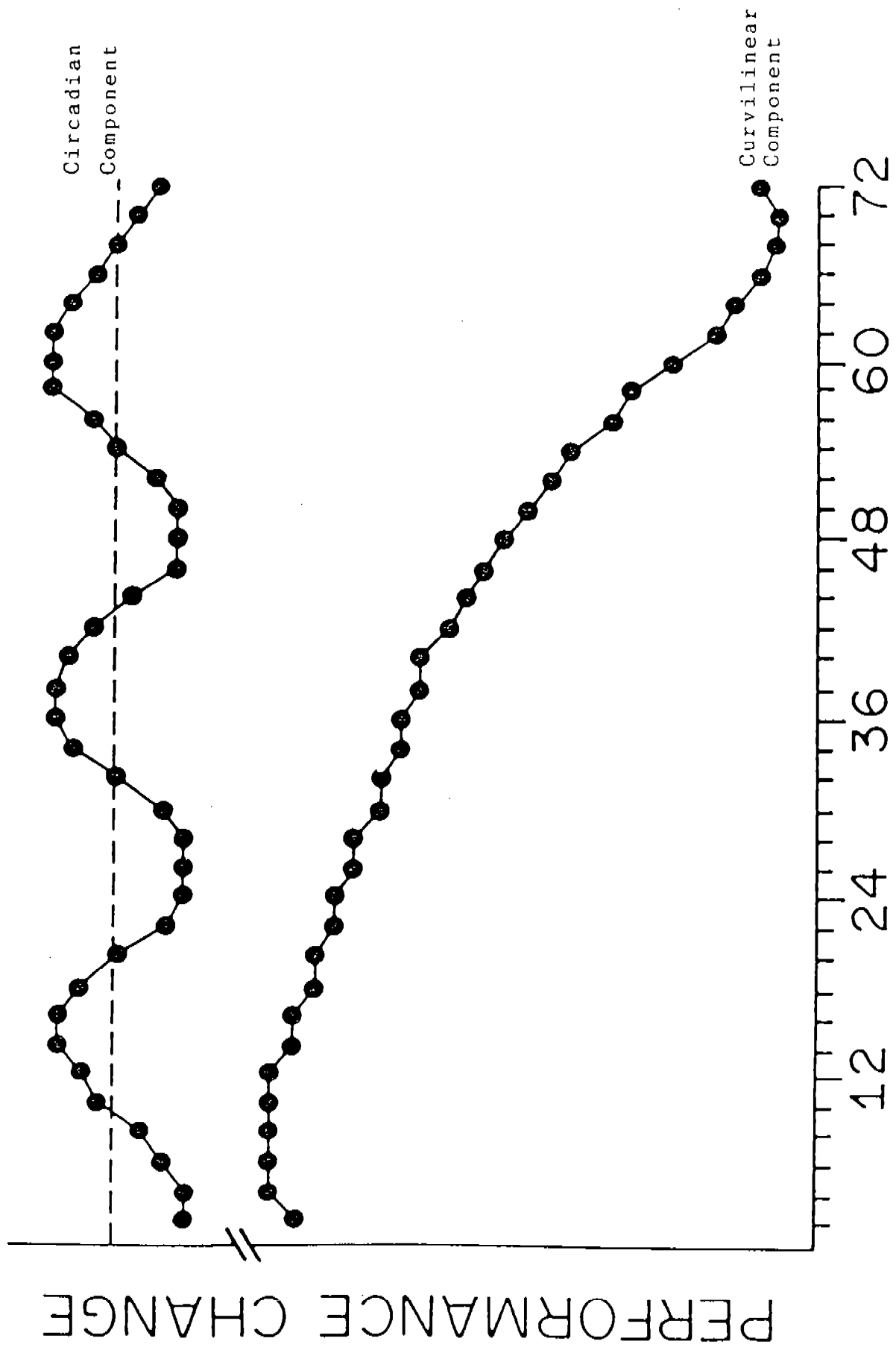


FIGURE 12

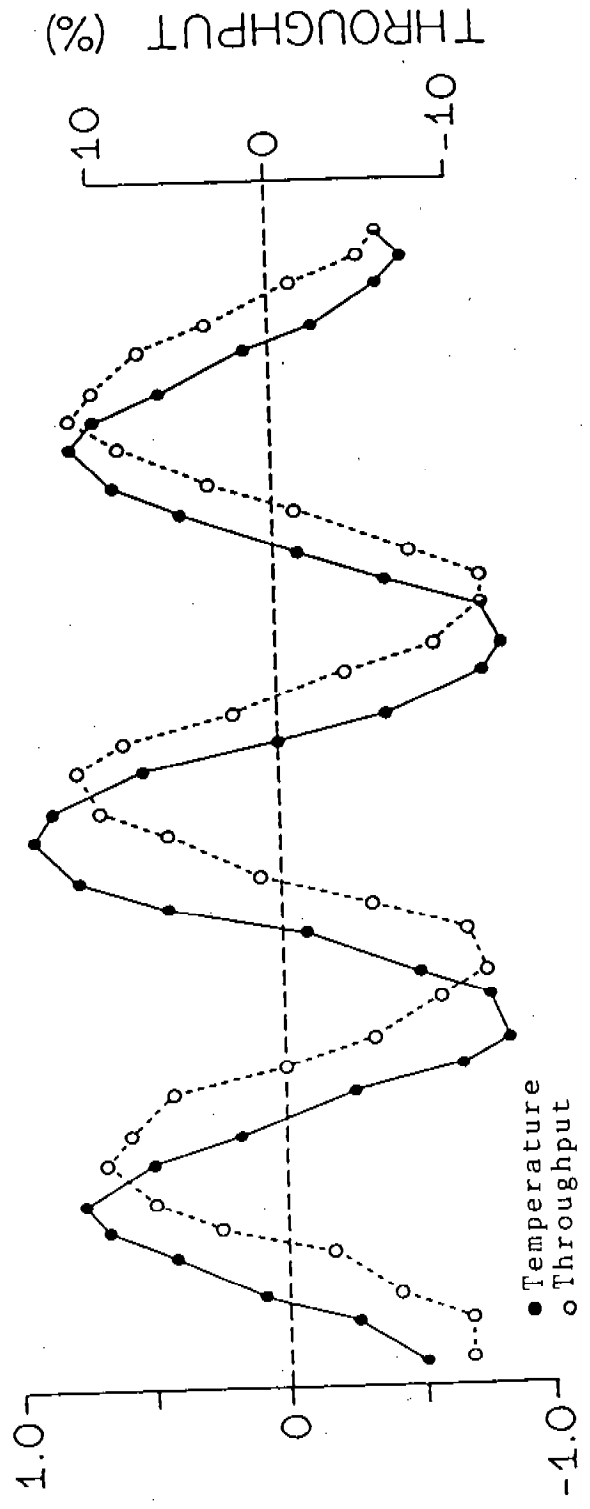
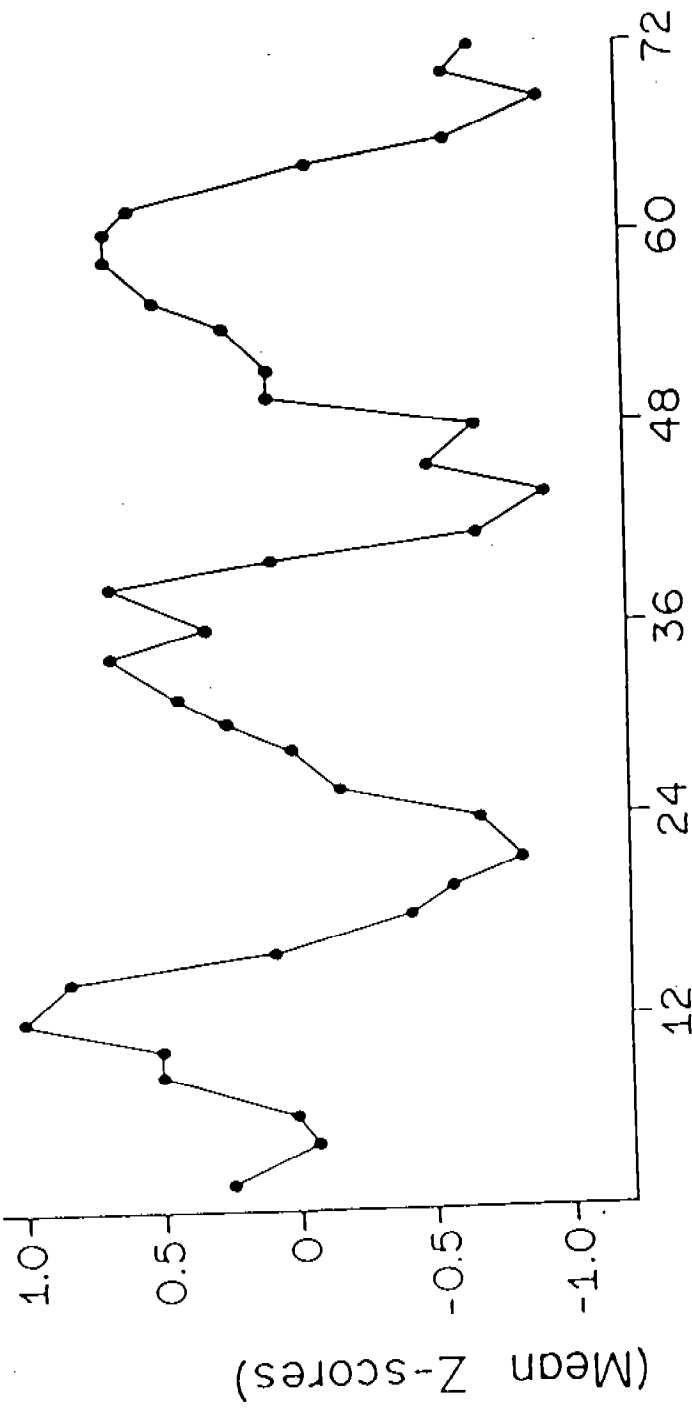


FIGURE 13

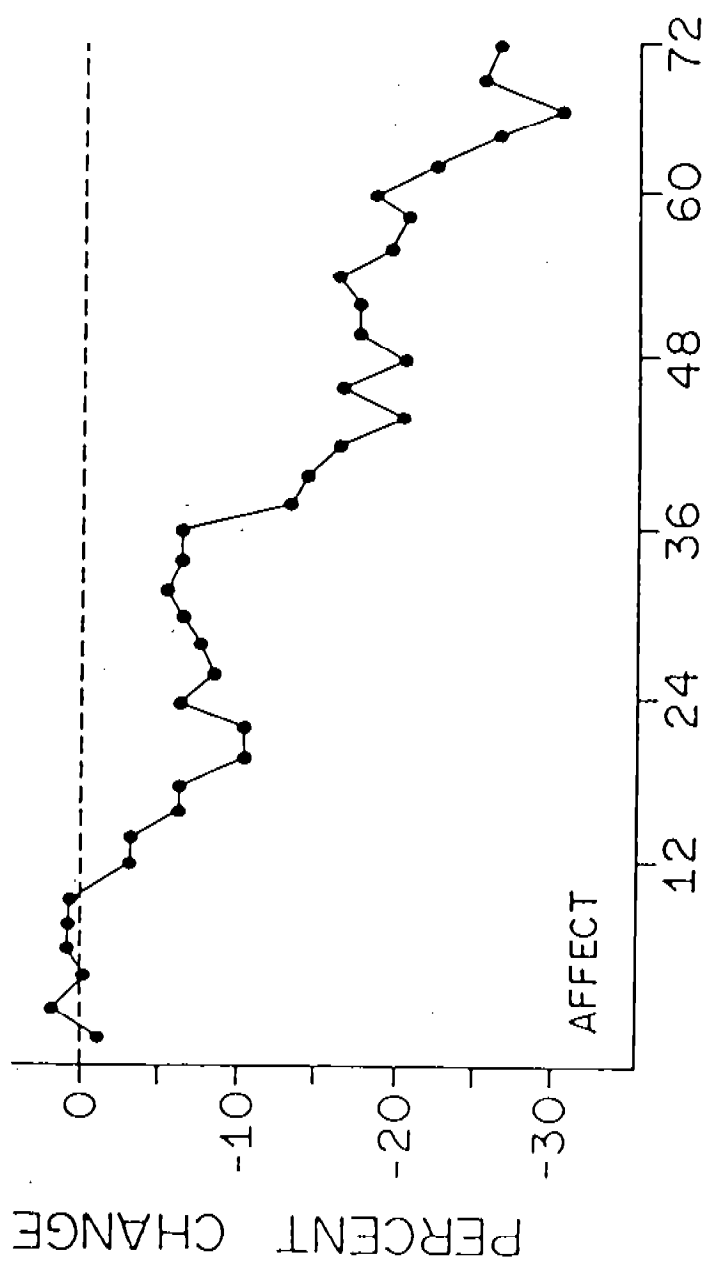
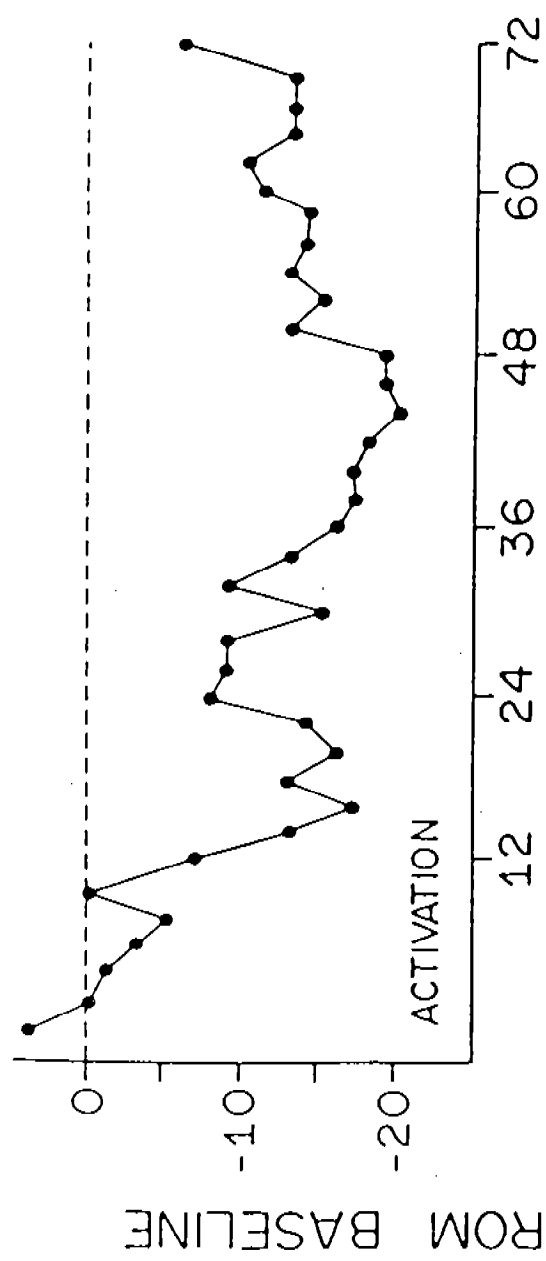


FIGURE 14

CIRCADIAN REMODULATE OF AFFECT/ACTIVATION

PLOT OF Y*X SYMBOL USED IS *

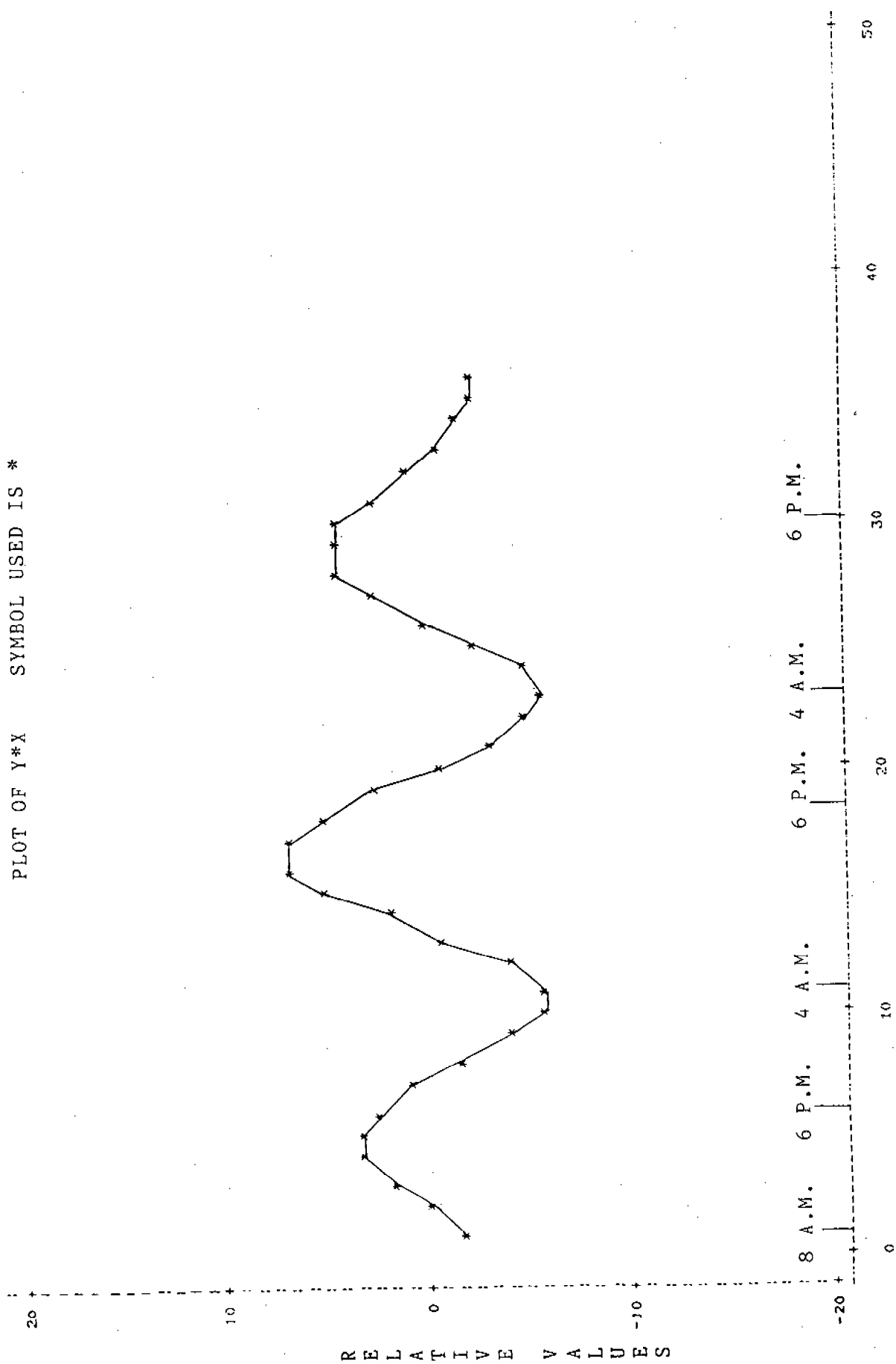


FIGURE 15

RELATIVE
MAGNITUDE

SUBJECT A

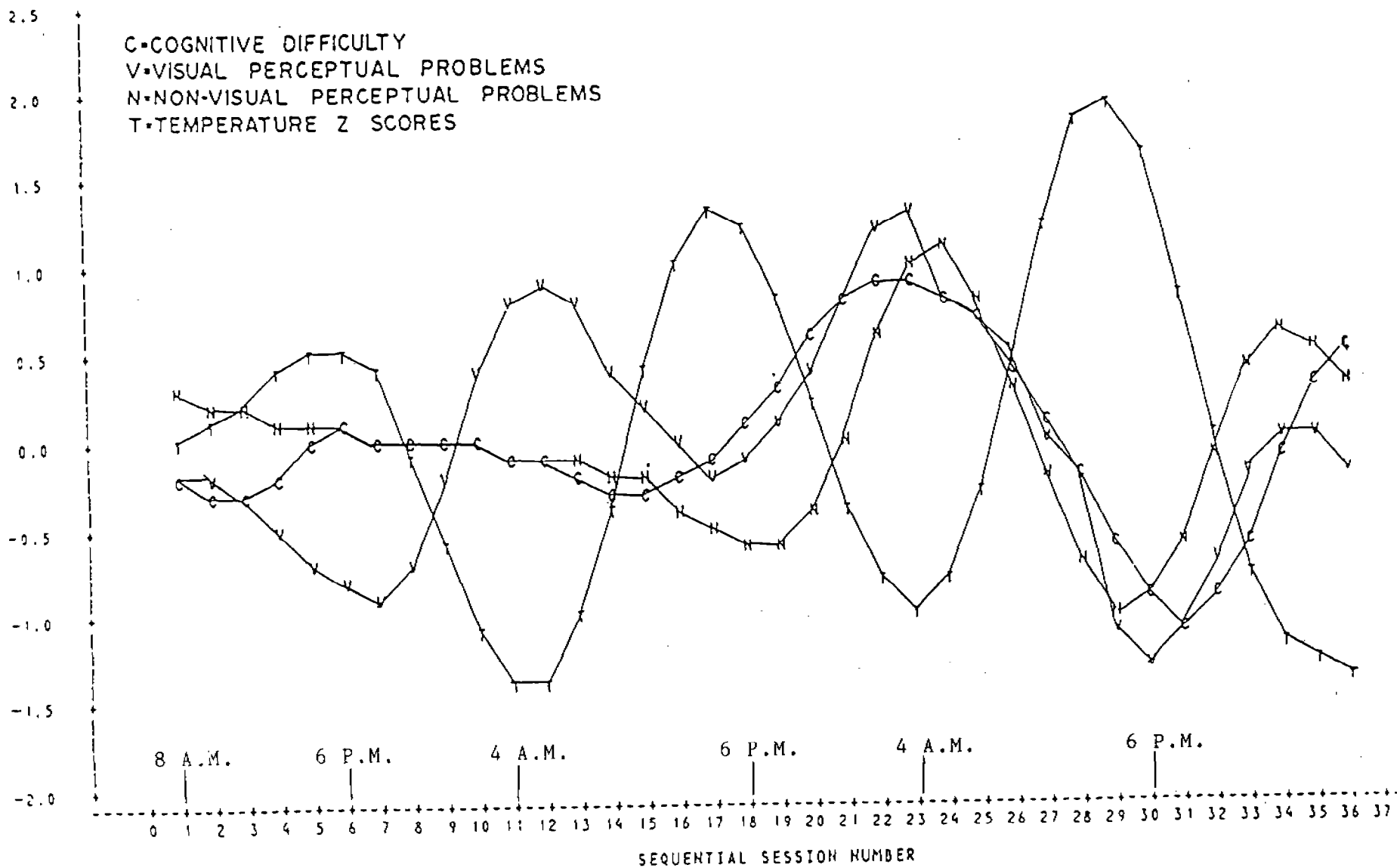


FIGURE 16

REFERENCES

1. Thorne, D.R., Genser, S. G., Sing, H. C., and Hegge, F. W.: Plumbing Human Performance Limits During 72 Hours of High Task Load. NATO Proceedings of the 24th DRG Seminar on The Human As A Limiting Element in Military Systems, Toronto, Canada, May, 1983: 1, 17-41.
2. Sing, H. C., Redmond, D. P., Hegge, F. W.: Multiple Complex Demodulation: A Method For Rhythmic Analysis of Physiological and Biological Data. IEEE Proceedings of the Fourth Annual Symposium on Computer Applications in Medical Care, Washington, D. C., 1980: 151-158.
3. Sing, H. C., Hegge, F. W., and Redmond, D. P.: Complex Demodulation - Technique and Application. Proceedings of the XVth International Conference of the International Society for Chronobiology, 1981, Minneapolis, Minnesota. New York, S. Karger Publishers, 1984.
4. Redmond, D. P., Sing, H. C., and Hegge, F. W.: Biological Time Series Analysis Using Complex Demodulation. In F. M. Brown and R. C. Graeber (eds) Rhythmic Aspects of Behavior. Hillsdale, N. J., Lawrence Erlbaum Associates, 1982, pp. 429-457.
5. Bloomfield, P.: Fourier Analysis of Time Series: An Introduction. New York, Wiley, 1976.
6. Blackman, R. B. and Tukey, J. W.: Modern Techniques of Power Spectrum Estimation. IEEE Trans Audio Electroacoust, AU-15: 55-66, June, 1967.
7. Thorne, D. R., Genser, S. G., Sing, H. C., and Hegge, F. W.: The Walter Reed Performance Assessment Battery. EPA Proceedings of Workshop on Neurotoxicity Testing in Human Populations, Raleigh, N. C., October, 1983.
8. Genser, S. G., Babkoff, H., Thorne, D. R., Sing, H. C., and Hegge, F. W.: Hallucination, Illusions, and Sleep Loss in Normals. (in preparation).
9. Pennington, R. H.: Introductory Computer Methods and Numerical Analysis. New York, Macmillan, 1965, pp. 404-411.

CYCLES OF SUICIDE

JOSEPH.M.ROTHBERG

WALTER REED ARMY INSTITUTE OF RESEARCH

WASHINGTON, DC 20307

Today's clinical presentation is the comparison of suicides in United States Army personnel, 1975-1982, and in the United States, 1972-1978. I intend to present our current epidemiological approach and point out some as-yet unresolved aspects of this work in order to solicit comments from this audience.

- * DETERMINE IF THERE ARE ARMY-SPECIFIC FACTORS AMONG SUICIDE IN ARMY PERSONNEL
- * H_0 : THERE ARE NO DIFFERENCES IN THE TIMING OF ARMY AND US SUICIDES.

Figure 1. Research goal and working hypothesis.

The goal of this observational study is to try to determine if there are meaningful fluctuations in the suicide data and to provide an analysis of the data base that identifies the correlates of any of these changes in the rates.

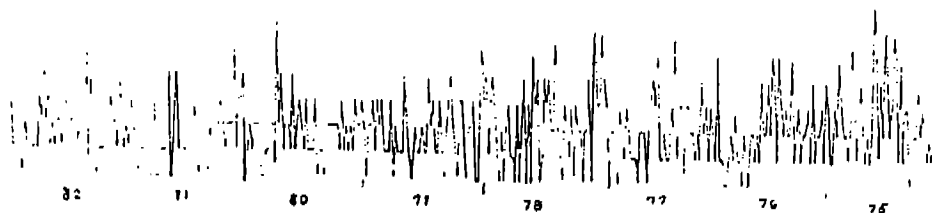


Figure 2. Weekly values of the numbers of suicides in United States Army personnel, 1975-1982 (note reversed scale).

Our Army data are the 834 suicides recorded during calendar years 1975 through 1982. These were 93% enlisted soldiers and 95% male. This is a sparse data set for the analysis of day-to-day trends since most of the 2922 days had no suicides. Figure 2 shows the number of suicides per week (the range is 0 to 7) from 1975 on the right edge thru 1982 on the left.

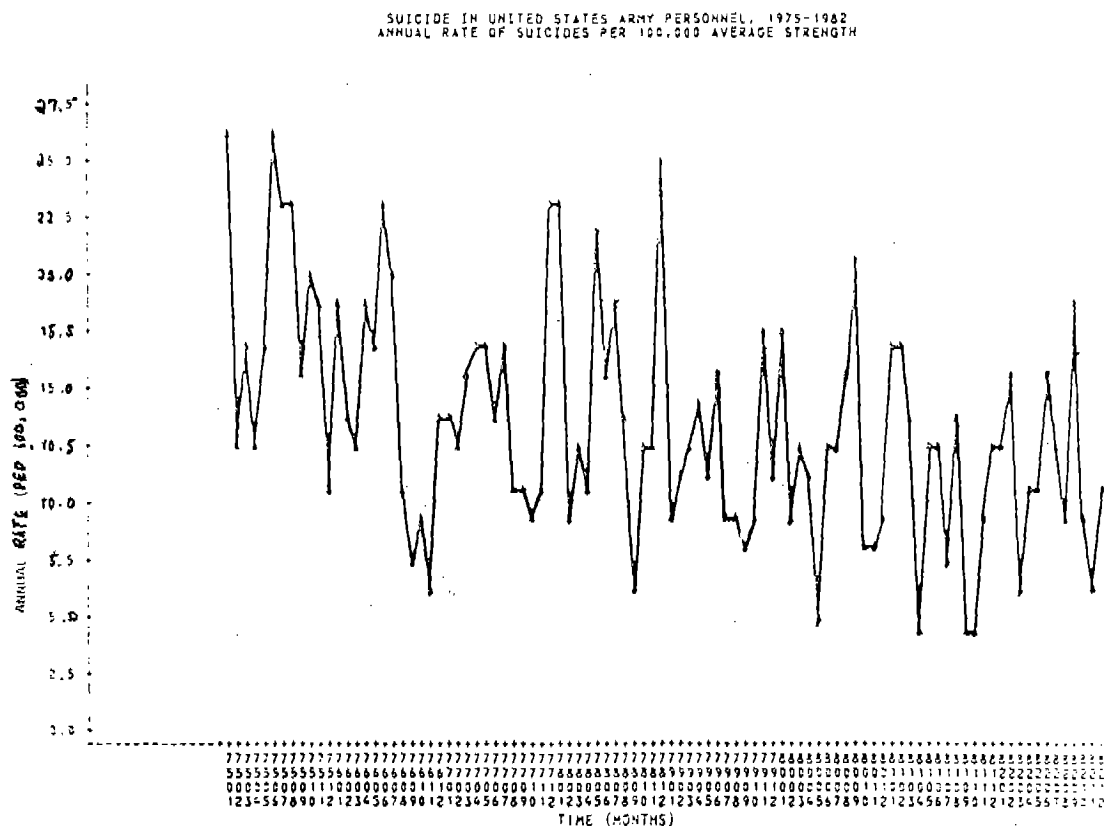


Figure 3. Monthly values of the annual rate of suicide (per 100,000) in United States Army personnel, 1975-1982.

Figure 3 shows the annual rate of suicides (per 100,000 average strength) in each month from January 1975 ('7501') through December 1982 ('8212').

As a starting point, we make the assumption that a population of soldiers will have the same suicide rate as the civilian population of the same age and sex. On a bi-annual basis, this turns out to be not entirely true. For each of the four bi-annual suicide reports (1,2,3,4), the male suicide rate is uniformly lower for the Army. For the females in the Army, their rate is not as reduced as is their male counterpart. Over all, the interpretation that the Army is a supportive social institution that protects against suicide is not contradicted.

Beyond this "zeroth level" comparison, the next set of questions were prompted by the paper of MacMahon (5) who reported on 185,887 suicides registered in the United States during 1972-1978. Her data presentation used the standard social units of time (week, month, year) and the lunar month. The percentage departure from the mean was plotted against the time span and cycles are apparent in the plots for all but the lunar month data. The Army data have been similarly arrayed and plotted along with the MacMahon data. The overlap of these two data sets is not complete since suicides by soldiers outside of the United States are only reported in the Army data. I will discuss these in order of increasing variability (distributing the same 834 cases into more intervals results in an increase in the variability).

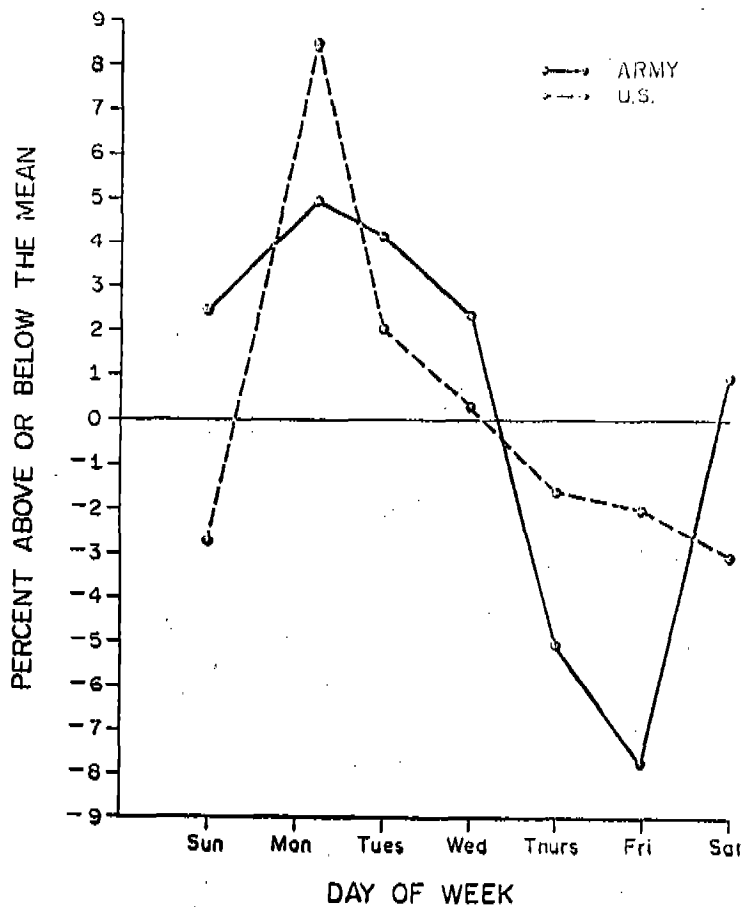


Figure 4. Deviation from the mean by day of the week for United States civilian population, 1972-1978 (U.S.) and United States Army, 1975-1982 (ARMY).

The day of the week data is shown in Figure 4. The two distributions appear to be quite similar. Both the Army and United States data show a Monday increase and a dip in the end of the week. For the United States, Saturday is the minimum while Friday is the minimum for the Army. The maximum departure from the mean is about the same for both data sets.

DAY OF WEEK

$$\text{CHI-SQ} = 1.24 \quad , \quad N = 7$$

$$P (1.24, 6) = 0.97 \quad , \quad \text{NOT SIG.}$$

Figure 5. Statistical test of day of week effect.

There is no significant difference between the two distributions on a chi-squared test.

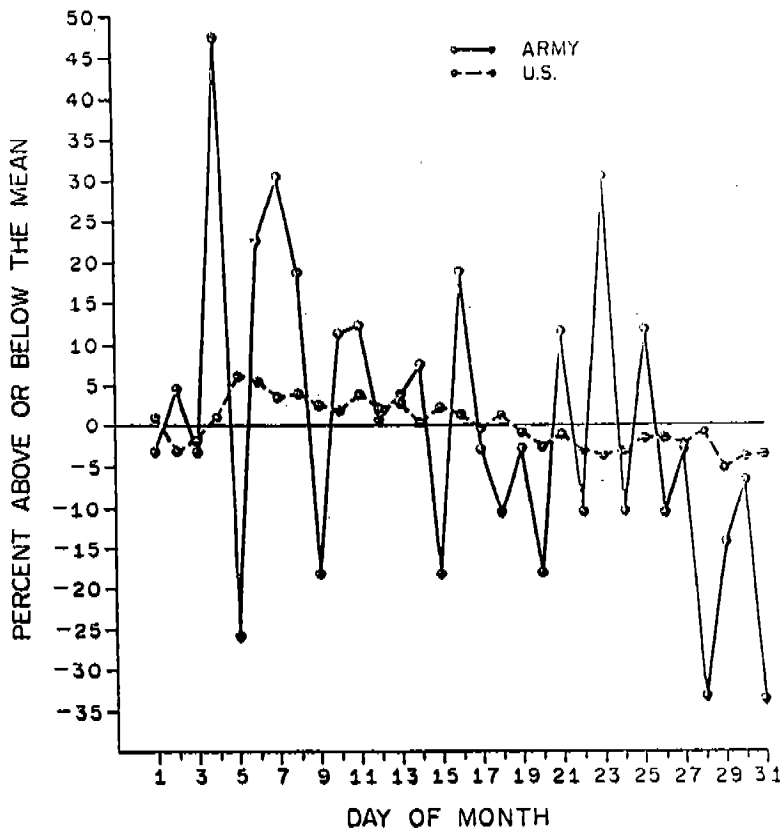


Figure 6. Deviation from the mean by month of the year for United States civilian population, 1972-1978 (U.S.) and United States Army, 1975-1982 (ARMY).

The month of year data are shown in Figure 6. Although both distributions have two relative peaks, they do not occur at the same time nor are they of the same amplitude. For the United States data, the peaks are less than 5% and occur in May and August/September. The Army has a peak in June that is almost 30% above the mean and a January peak is almost 25% above the mean.

MONTH OF YEAR

$$\text{CHI-SQ} = 22.56, \quad N = 12$$

$$P(22.56, 11) = 0.02, \quad \text{SIG.}$$

Figure 7. Statistical test of month of year effect.

The probability that these distributions are the same is only 0.02. Some military reassignments to new posts occur at about those times. The stress of relocation is a plausible precipitant of suicide.

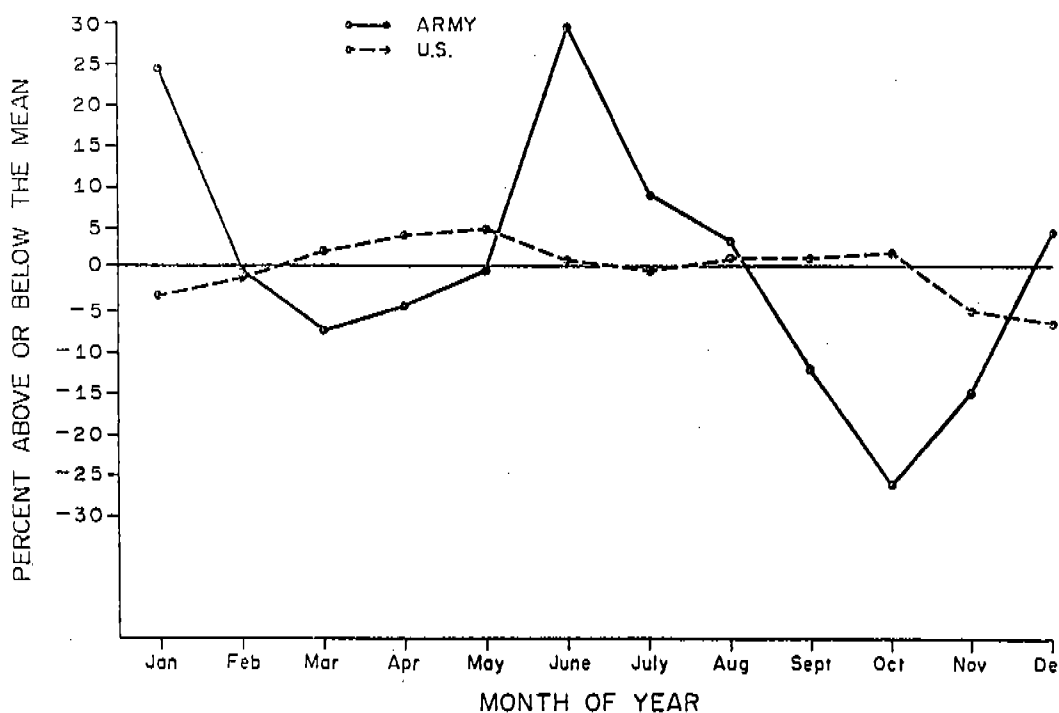


Figure 8. Deviation from the mean by day of the month for United States civilian population, 1972-1978 (U.S.) and United States Army, 1975-1982 (ARMY)..

The day of month data are shown in Figure 8. The United States data shows a peak on the fifth of the month followed by decreasing values until the end of the month. The Army data have a great deal of variability but, using a five day sliding average (not shown) there appears to be a set of peaks early in the month (on the 4th, 6/7th, and 10th) and a peak late in the month (on the 22nd) and a dip at the end of the month (on the 28th).

DAY OF MONTH

$$\text{CHI-SQ} = 26.03, \quad N = 31$$

$$P (26.03, 30) = 0.67, \quad \text{NOT SIG.}$$

Figure 9. Statistical test of day of month effect.

There is no significant difference between the distributions using a chi-squared test. Pay day in the Army is the last working day of the month and some of the suicides may be due to financial problems that become apparent close to pay day and the first-of-the-month bills.

What we have done in discussing these figures was to average the eight years of data assuming that there are cycles of psychosocial events occurring at specified times which drive these suicides. The increased rates at the start of the week, the start of the month and the start (and middle) of the year lend support to the assumption that there are cycles.

The question of cycles within the Army suicide data was looked at directly but only briefly. We did a spectral decomposition of the daily suicide counts using the SAS procedure SPECTRA.

SAS

PROCEDURE SPECTRA WHITETEST

2916 DAYS , 831 SUICIDES

H₀: THE LARGEST OBSERVED PERIODOGRAM
ORDINATE IS THE LARGEST IN A
SIMILARLY SIZED RANDOM SAMPLE

H'₀: THE FREQUENCY SPECTRUM IS NOT
DIFFERENT FROM WHITE NOISE.

FISHER'S KAPPA = 7.66, 1457 D.F.

P (7.66, 1457) > 0.10 , NOT SIG.

Figure 10. Statistical test of periodogram randomness.

Since the fast fourier transform algorithm of that procedure requires that the number of data points have a largest prime divisor less than or equal to 23, the analysis was done with the first 2916 days. The null hypothesis that the largest observed periodogram ordinate is the largest in a similarly sized random sample was tested with Fisher's Kappa. The value of 7.66 with an n of 1457 two-degree-of-freedom periodogram ordinates has a $p > 0.1$. With that negative result, it appears that any search for further structure within the Army suicide data would be inappropriate.

The inability to proceed further with the analysis of the Army suicides for cycles in a direct fashion shouldn't interfere with having clever ideas about the cyclic properties of the United States data and then testing if the Army data looks like the United States data. And it is at this point, needing some clever ideas, that I solicit the audience to suggest ways to look at this relatively small but important data set.

REFERENCES

1. Suicide in United States Army Personnel, 1975-1976.
Datel, W.E., and Johnson, A.W., MIL MED 144(4):239-244, 1979.
2. Suicide in United States Army Personnel, 1977-1978.
Datel, W.E., Jones, F.D., and Esposito, M.E., MIL MED 146(6):
387-392, 1981.
3. Suicide in United States Army Personnel, 1979-1980.
Datel, W.E., and Jones, F.D., MIL MED 147(10): 843-847, 1982.
4. Suicide in United States Army Personnel, 1981-1982.
Rothberg, J.M., Rock, N.I., and Jones, F.D., MIL MED (in press).
5. Short-Term temporal cycles in the Frequency of Suicide,
United States, 1972-1978. MacMahon, K., AM J EPIDEMIOL
117:744-750, 1983.

EVALUATION OF OPTICAL DATA COLLECTION INSTRUMENTATION IN THE DESERT ENVIRONMENT

Robert A. Dragon
National Range, Data Collection Division
US Army White Sands Missile Range
White Sands Missile Range, NM 88002

ABSTRACT The integration of new technologies such as video systems in place of current high-speed film cameras is discussed. For a great percentage of daytime activity the desert atmosphere is shown to be a limiting factor for the collection of visual data. The atmosphere and instrument focal lengths are hypothesized to be major considerations for instrument design both with film and video systems. An experiment using existing data and analysis of variance is suggested to evaluate the hypothesis.

INTRODUCTION The data collection task at White Sands Missile Range (WSMR) often relies on a photographic record consisting of accurate images of test projectiles. Photography has been the common method of securing these records, usually through the use of tracking telescopes, cinetheodolites, television, and fixed cameras. As new technologies become available, it is natural to expect them to be rapidly integrated into the full complement of existing optical instrumentation.

New technologies and instrument performance have been important considerations at WSMR for more than thirty-five years. Consideration of the atmospheric environment and its interaction with the optical system has always been considered important.^{1,2} However, until the midninteen seventies field implementation of video systems was not entirely practical because of low frame rate. During this period projected video requirements were discussed in detail.³

Various articles have recently appeared comparing the advantages of video over photographic systems.^{4,5} Although these discussions are both timely and appropriate, many researchers fail to include the actual field conditions as a significant factor which contributes to image quality.

This is a 'best case' analysis. That is, only some of the factors which may degrade optical system performance have been considered. Other factors such as mechanical vibration and photographic processing are outside the scope of this paper.

THE ATMOSPHERE The desert atmosphere can be one of the best and well behaved components of any optical system. However, meteorological deterioration can be significant, especially in the daytime. For long distances (and even short distances if one is viewing close to the ground) one can expect some sort of image degradation. Local ground heating, wind, and dust can seriously degrade images at both low and high angles of observation. Excellent night seeing resolution is about one arc-second ("). Poorer seeing resolution is often the case over most of the daytime southwest desert.

THE RECORDING MEDIUM The effect of the recording medium on the recorded image is important. Two common recording media, photographic film and video will be compared. In this analysis Ektachrome film with a high contrast resolution of 70 line pair per millimeter (lp/mm) is used. If the entire tape-playback system is considered, the resolution of most current video systems is about 17 lp/mm.

When atmospheric seeing is degraded each point is imaged as a much larger point. The image size is given by:

$$\text{Image size} = \frac{\text{lens}}{\text{focal length}} \times \text{angular resolution} \quad [1]$$

$$\times 4.85 \times 10^{-6} \text{ arc-sec}$$

where 1 arc-second = 4.85×10^{-6} radians.

The following seeing resolutions can thus be translated into linear resolutions and lp/mm at the photographic or video receiver as follows.

Table 1

Seeing	Image size	Resolution	System
1 arc-sec	12 μm	81 lp/mm	100-inch focal length
5 arc-sec	62 μm	16 lp/mm	
10 arc-sec	124 μm	8 lp/mm	
1 arc-sec	24 μm	40 lp/mm	200-inch focal length
5 arc-sec	124 μm	8 lp/mm	
10 arc-sec	248 μm	4 lp/mm	

System resolution is computed by the following

$$1/R_T = 1/R_1 + 1/R_2 + \dots 1/R_N \quad (\text{Ref 6}) \quad [2]$$

Where $R_1, R_2, \dots, R_N$ are the component resolutions and R_T is the system resolution.

This relationship between atmospheric seeing, system focal length, and resolution can be shown graphically by the figure below.

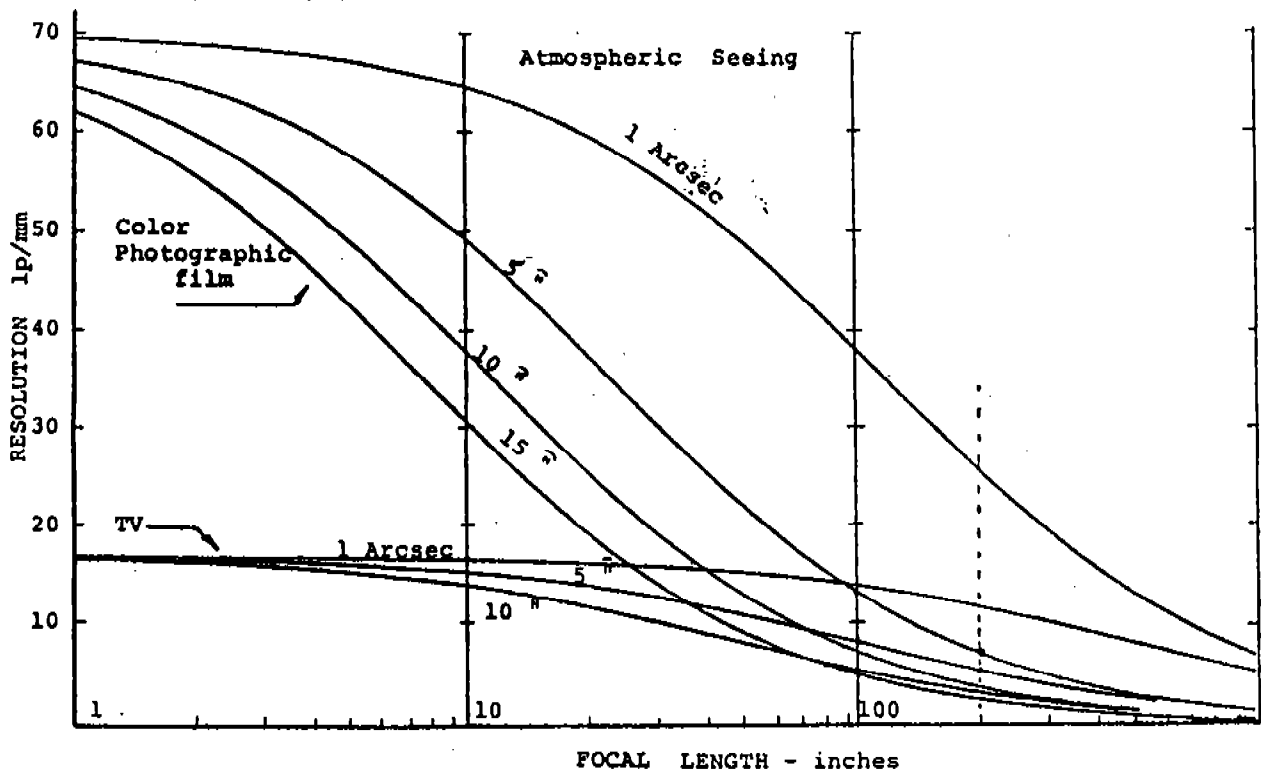


Figure 1

ANALYSIS If this hypothesis (i.e. the above relationship) is to be tested, large amounts of data regarding the instruments and quality of optical data will be required. Fortunately this data is currently available in the form of film/video analysis records for each test mission. These records cover about one year of previous testing. The following have been selected as relevant variables to determine any relationship between record quality and any of the variables. They are:

1. Recording Medium (film or TV)
2. Lens Focal length
3. Weather
4. Time of Day
5. Test name
6. Equipment operator
7. Instrument Site
8. Instrument number

Analysis of variance methods are proposed for the analysis of this data.

	FILM	VIDEO
<u>LENS FOCAL LENGTH</u>		
50 inches		
50 100 inches		
100 inches		
200 inches		
<u>WEATHER</u>		
CLEAR/GOOD		
WIND/DUST		
WIND/NO DUST		
CLOUDY		
<u>TELESCOPE OPERATOR</u>		
NUMBER 1,2,... etc.		
<u>TELESCOPE NUMBER</u>		
NUMBER 851,852,... ect.		
<u>TIME OF DAY</u>		
HOUR 06:00, 07:00,...etc.		

The data is a record of image quality. Although somewhat subjective, sufficient records should show the hypothesized relationships clearly.

ANALYSIS OF DATA

It is proposed to analyze the data by use of multivariate Analysis of Variance. An excellent treatment of this subject arranged for use on computer is given by Jeremy D. Finn¹⁰.

CONCLUSIONS

With little additional input sufficient data can be extracted from existing data records to show relationships between the use of film or video and other variables. Advantages (or disadvantages) between the use of photographic film (other than cost) should be clearly shown.

REFERENCES

1. Sixby, S.R., Long Focal Length Ballistic Camera Study, Perkin Elmer Electro-Optic Division, Norwalk, Conn., October 1961
2. Fahay, Thomas P., Astro-optical Tracker Study, APGC-TR-60-42, perkin Elmer Corp., Norwalk, Conn., March 1960.
3. Newton, Henry L. and Davidson, Charles W., High Resolution Television, Data Systems Technical Memorandum 74-8, White Sands Missile Range, New Mexico, July 1974.
4. Scrivener, C.B. Jr., Film to Video Conversion Study, Pan Am World Services, Inc. and RCA Corp., Eastern Space and Missile Center, Air Force Systems Command, Patrick Air Force Base, Florida, December 1982.
5. Baker, Ralph L., An Assessment of the Feasibility of Converting Cinetheodolites to Videothodolites at White Sands Missile Range, Baker Electronics Systems, Tucson, Arizona, October 1980.
6. Arnold, C.R., Rolls, P.J., Stewart, C.J., Applied Photography, The Focal Press, London, 1971.
7. Armore, Sidney J., Elementary Statistics and Decision Making, Charles Merrill Co., Columbus, Ohio, 1973.
8. Cooper, B.E., Statistics for Experimentalists, Pergamon Press, Oxford, 1969.
9. Volk, William, Applied Statistics for Engineers, McGraw-Hill, New York, 1969.
10. Finn, Jeremy D., Multivariate Analysis of Variance and Covariance, Statistical Methods for Digital Computers, Ed. Kurt Enslein, Anthony Ralston, Herbert S. Wilf, Wiley, New York, 1977.

A TYPE OF CORRELATED DATA IN OPERATIONAL TESTING

Ellen Hertz
U.S. Army Operational Test & Evaluation Agency
Falls Church, VA

ABSTRACT. During a portion of a test, N gunners fired two rounds apiece. The overall proportion of hits on first rounds was very close to the overall proportion of hits on second round shots. However, an individual gunner's performance on his second shot was positively correlated with his performance on the first round.

The parameter of interest was p , the probability of hit using the firing device. The proportion of hits among the $2N$ shots was the natural point estimate of p . However, in calculating interval estimates for p at a given confidence level, or tests of hypothesis of the form $p \geq p_0$ at a given significance level, the situation became more subtle. Since the first round outcome did not deterministically predict the second round outcome, we clearly had more information than just the N first round shots. On the other hand, the assumption that we had $2N$ independent trials was not justified.

In this paper, a model is proposed for the analysis of this and similar situations. This model generalizes the "two round" case and considers data in blocks when the observations within blocks are not independent.

I. INTRODUCTION. During a portion of the test of a firing device, each gunner fired a volley consisting of two rounds. The outcome of each round was either hit (H) or miss (M) and one of the purposes of the test was to draw inferences about p , the probability of hit.

The following table depicts a typical segment of the results:

Gunner										
Rnd	1	2	3	4	5	6	7	8	9	10
1	H	H	M	M	H	M	H	H	M	H
2	H	H	H	M	H	M	H	M	M	H

Here, the overall proportion of hits on a first round is .6 and the overall proportion of hits on a second round is also .6. The probability of hit on a first round appears to be the same as the probability of hit on a second round, so the overall proportion of hits is an unbiased point estimate of p . However, the conditional probability of hit on a second round after having scored a hit on the first round of the volley is $5/6$ which is greater than .6. In other words, performance on the second round is not independent of performance on the first round. Suppose n volleys were fired. We do not have $2n$ independent rounds. On the other hand, since the outcome on the first round did not predict the outcome on the second round deterministically, we have more information than just the n first round shots. The problem is to calculate confidence intervals and tests of hypotheses about p that reflect our true amount of knowledge realistically.

II. THE MODEL. n players are selected at random. The probability of hit for a player comes from a distribution with mean p and unknown variance σ^2 . Then $P_1, \dots, P_n$, the players' hit probabilities, are independent and identically distributed random variables with mean p .

The i 'th player fires k_i shots, $k_i \geq 1, i=1, \dots, n$. The data is $\{X_{ij}: i=1, \dots, n, j=1, \dots, k_i\}$ where $X_{ij}=1$ if the i 'th player scored a hit on the j 'th trial and 0 otherwise. If $i \neq j$ then X_{ir} and X_{js} are independent. X_{ir} and X_{is} are correlated but are conditionally independent Bernoulli variables with parameter p_i given $\{P_i = p_i\}$.

III. THE TEST STATISTIC. Set $G_i = \sum_{j=1}^{k_i} X_{ij}$, $i=1, \dots, n$ and let $T = \frac{1}{n} \sum_{i=1}^n (G_i/k_i)$. Then, using the law of conditional expectation, $E(G_i) = EE(G_i|P_i) = E(k_i P_i) = k_i p$ so that T is an unbiased estimate of p .

$$\begin{aligned} E(G_i^2) &= EE\left(\sum_{j=1}^{k_i} X_{ij}^2 + \sum_{j \neq r} X_{ij} X_{ir} \mid P_i\right) = \\ &E(k_i P_i + k_i(k_i-1)P_i^2) = k_i p + k_i(k_i-1)(p^2 + \sigma^2) \text{ so that} \\ \text{Var}(G_i) &= k_i(p-p^2) + \sigma^2 k_i(p-p^2) + \sigma^2(k_i^2 - k_i). \end{aligned} \quad (1)$$

If we set $A = \sum 1/k_i$ then

$$\text{Var}(T) = (A(p-p^2) + \sigma^2(n-A))/n^2 \quad (2)$$

To utilize T as a test statistic, it is necessary to estimate $\text{Var}(T)$.

The following lemma is easy to verify: If $Y_1, \dots, Y_n$ are independent with a common mean and $\text{Var}(Y_i) = \sigma_i^2$, $i=1, \dots, n$ then $E \sum_{i=1}^n (Y_i - \bar{Y})^2 =$

$(n-1)/n \sum_{i=1}^n \sigma_i^2$ Applying the lemma with $Y_i = G_i/k_i$ and using (1),

$$E \sum_{i=1}^n (G_i/k_i - T)^2 = ((n-1)/n)(A(p-p^2) + \sigma^2(n-A)). \quad (3)$$

Letting $D = \sum (G_i/k_i - T)^2$, it follows from (2) and (3)

that $D/(n(n-1))$ is an unbiased estimate of $\text{Var } T$. The statistic that is proposed is, then, T/E where $E = \sqrt{D/n(n-1)}$. If $P[U \leq x] = 1 - \alpha/2$ for U standard normal then $T - Ex \leq p \leq T + Ex$ is an approximate $1 - \alpha$ confidence interval for p . Another application would be to test the hypothesis $H_0: p \geq .9$ vs. $H_1: p < .9$ using the rejection criterion $(T - .9)/E \leq -x$ to achieve a significance level of approximately $\alpha/2$.

IV. A REFINEMENT. If $C_1, \dots, C_n$ are any real numbers such that

$\sum_{i=1}^n C_i k_i = 1$ then $T^* \equiv \sum C_i G_i$ is an unbiased estimate of p . The choice

of $C_i = 1/(nk_i)$ was made to facilitate estimating the variance of T^* .

This corresponds to weighting each player equally. Another possibility would be $C_i = 1/N$, $N = \sum k_i$, i.e. weighing each shot equally. Using Lagrange multipliers to minimize $\sum C_i^2 \text{Var } G_i$ subject to the condition

$\sum C_i k_i = 1$ yields the result $C_i = K/(p - p^2 + \sigma^2(k_i - 1))$ where K is a constant of proportionality.

V. A SIMULATION. Since normal approximation was used, a simulation was run to test the accuracy of this method. A situation was considered in which four players were selected. Their probabilities of success were distributed uniformly on $[.5, 1]$ so that the overall probability of success was .75. Each player fired 5 shots. 95% confidence intervals were constructed using both the proposed statistic and using $(4)T \pm 1.96\sqrt{T(1-T)/N}$ i.e. neglecting the heterogeneity of the players. The program calculated the proportion of times the confidence interval contained .75, the true value of p .

For three runs, the results were .97, .96 and .97 for the proposed interval and .81, .77 and .78 using (4).

APPENDIX - SIMULATION PROGRAM

```

5  X=0:Y=0
10 DIM P(4), X(4,5), G(4)
15 CNT=0
20 FOR I=1 to 4

30 P(I)=.5*RND(1)+.5
40 FOR J=1 to 5

50 X(I,J)=0

60 H=RND(1)

70 IF H <=P(I) THEN X (I, J) =1

80 NEXT J: NEXT I
85 T=0
90 FOR I= 1 to 4
100 G(I)=0
110 FOR J=1 to 5
120 G(I)=G(I)+X(I,J) : NEXT J
130 T=T+G(I) : NEXT I

140 T=T/20
150 D=0
160 FOR I=1 to 4 : D=D+(G(I)/4-T) ^2

170 NEXT I
180 E=SQR (D/12)

200 IF ABS (T-.75) <=1.96*E THEN X=X+1

210 IF ABS (T-.75) <=1.96*SQR (T*(1-T)/20) THEN Y=Y+1

220 CNT=CNT+1

230 IF CNT <500 THEN 20

240 PRINT "XBAR="; X/500; "YBAR=";Y/500

250 END

```


A Simulation Process for Determining Reliability of Cyclic Random Loaded Structures

D. Neal, W. Matthews and T. DeAngelis
Army Materials and Mechanics Research Center

Abstract

A unique application of the Monte Carlo method was developed for determining reliability vs. cycles to failure of the M60 tank torsion bar. In applying the method, material torsional fatigue and spectrum loads were modelled such that variability in the functional parameters and operational loads were represented. Random torsional displacement values obtained from the amplitude displacement distributions applied to the fatigue equations resulted in an exponential distribution for cycles to failure of the in service bar. The number of simulations in the Monte Carlo process was determined from a convergence criteria involving stability of the third and fourth moments of the cycles to failure distribution.

Reliability vs. bar life computations indicated a negligible amount of life after flaw initiation. Assuming a design change involving a twenty percent reduction in bar stresses increased the life estimates by a factor of three. An increase in reliability can also be realized if computations are made by assuming a bar has been in operation for a specified number of cycles. A comparison of minimum life (ninety nine percent probability of survival) between predicted and in service results showed excellent agreements (less than eight percent difference).

Introduction

The current need for establishing reliability of various components and systems for U.S. Army weapon vehicles is being realized. The consequences of over- or under design are often reflected in either premature failure or excessive costs and poor performance due to excessive weight. The mean life estimates used as a criteria for defining acceptability of cyclic loaded component will often provide a false sense of security regarding its capability. The application of higher strength ferrous materials or the less conventional structural materials such as composites and ceramics will often result in premature failure because of the inability to recognize the inherent variability of the materials strength.

The objective of this paper is to determine a methodology which will circumvent the present deterministic approach used in establishing an acceptable design for cyclic random loaded structure. Instead of analyzing the worst case situation related to the spectrum loads, S/N curve, or crack propagation laws, the authors introduce a method which simulates the variability in loading and materials capability. Use of this methodology eliminates the over (worst case) or under design (mean life) situation by introducing a probabilistic design criteria. Recognition of the reliability values as a function of the life cycles of operation can provide the opportunity for selecting a specified life value corresponding to the probability estimate. The remaining component life can then be determined as related to its probability number.

The recommended ASTM procedure for determining acceptable design, involves establishing a lower confidence 3 Standard Deviation bound on the S/N Curve then selecting cycles to failure from the bounded curve consistent with predetermined maximum stress obtained from the spectrum load results. This procedure can often result in an over design situation since the maximum load may rarely occur in addition to the fact there is a small chance that the lower S/N Curve bound is representative of the True S/N Curve.

The Monte Carlo process used in predicting life time versus reliability of the M60 torsion bars had a prior application in a report by (1). Conceptually, this method is quite simple, requiring modelling of the spectrum loads and the material fatigue life with respect to crack propagation or stress/cycles to failure.

Amplitude Displacement Model

In figure 1, a schematic of the torsion bar in the M60 Tanks is shown. The amplitude distributions of three bars from tests conducted at Aberdeen Proving Grounds (APG) is shown in figure 2. Positive and negative angular displacements of the bars as function of tank travel are shown in figure 2a. In figure 2b the amplitude distributions are listed in a manner describing percent time less than by a plus sign (+) and percent time greater than by a minus sign (-), (eg. 25% level equals a -75% level. The + peak represents maximum angular displacement under load, the negative peak is maximum unloaded angular measure. In order to eliminate considering positive and negative peak values in figure 2a for determining angular displacements in the cyclic loading process, the angular displacement is defined as follows,

$$\Delta \theta = \theta + |\theta^-|$$

where θ^- = maximum negative angular displacement (1)

θ = displacement from figure 2b

$\Delta \theta$ represents the adjusted angular displacement

The Beta distribution provided the best representation of the skewed amplitude distribution. The dampening effects that occurred under load resulting from a stop used in preventing further angular twist of the bar producing a highly skewed discrete cumulative probability values. The Beta function is defined as:

$$f(\Delta \theta) = \frac{\Gamma(P+Q)}{\Gamma(P)\Gamma(Q)} (\Delta \theta)^{P-1} (1-\Delta \theta)^{Q-1}$$

and $0 \leq \Delta \theta \leq 1$ $P, Q > 0$ (2)

The P and Q values are selected in a manner that provides the best Probability Density Function (PDF) for representing the data. Figure 3 describes a typical distribution and Table 1 shows the excellent correlation between predicted (Beta representation) and actual test results. Angles less than 20° represent stresses sufficiently low that infinite torsion bar life could be expected, therefore, a good representation below this angle is not essential.

Crack Growth Law For Estimating Torsion Bar Life

Initial efforts in applying the Monte Carlo Method for determining reliability vs cycles to failure of the torsion bar involved using the crack propagation laws. The da/dN relationships for materials metallurgically similar to the specified material were obtained from (2), (3), and (4) and is shown in figure 4. The dry air results made available by Barsom (4) provided the most representative estimates of crack growth vs stress intensity (ΔK) described in figure 4 since the torsion bar is protected from the environment. From the basic da/dN relationship, N cycles to failure as a function of crack growth, angular displacement and the geometry of the region where the crack initiates in the bar, may be obtained from the following relationships:

$$N = \int_{c_1}^{c_f} \frac{dc}{.66 \times 10^{-8} \Delta K^{2.25}} \quad (3)$$

$$\text{where } \Delta K = A_j \Delta \theta \sqrt{\pi C}$$

$$\text{and } A_1 = 4.91 \text{ (Key Way)}$$

$$A_2 = 3.29 \text{ (Other Spline Regions)}$$

$$A_3 = 3.26 \text{ (Shaft Section)}$$

Note, a percent reduction in A_j 's will provide a decrease in the stresses in the specific region of the torsion bar. The C_i and the C_f parameters are initial and critical crack size respectively. The C_f is obtained from critical stress intensity value K_{Ic} for the material considered. The angular displacement of the bar can also be represented by the equivalent stress value τ as

$$\text{Max } \tau = rG (\Delta \theta) / L$$

$$r = \text{radius of shaft}$$

$$G = \text{torsional modulus}$$

$$\Delta \theta = \text{max. allowed angle}$$

$$L = \text{length of torsion bar}$$

(4)

The Monte Carlo Process

(A) Crack Propagation Analysis

A schematic of the process is outlined in figure 5 for determination of frequency of occurrence vs. cycles to failure of the torsion bar using the crack propagation law. An assumed normal distribution is used to represent variability in the A_j , C_i , and C_f parameters. A coefficient of variation (C.V.) defined as

$$\text{C.V.} = \frac{\text{S.D.}}{\text{mean}}$$

(5)

establishes the standard deviation S.D. for the corresponding known mean value (eg $\bar{C}_i$ for initial crack size). C.V. values of 5, 10 and 15 percent were considered in developing the distributions in order to examine the effects of variability (inherent errors in measurements, flaw size assumption or the stress analysis) in the parameters. By selecting the above C.V.'s a sensitivity analysis can be developed, thereby providing a method for recognizing the importance of the parameters as related to cycles to failure number. The Beta distribution as shown in figure 5 has been previously defined in equation (2).

The random numbers used in the Monte Carlo process are obtained from solving for X in

$$\int_{-\infty}^X f_i dX = R$$

(6)

where R is a uniform random number and f_1 corresponds to the desired type of frequency distribution for the parameter. A probability density function for the N cycles to failure can be obtained by randomly selecting from C_i , C_f , A_i , and $\Delta\theta$ distributions of discrete sets of numbers and substituting them into equation 3. Note, there should be an equal amount of random numbers for each parameter to have proper amount of numbers for the N distribution.

(B) S/N Curve Analysis

Torsional bar life expectancy was obtained using the Monte Carlo process applied to the S/N Curve relationship. The procedure provided a method for obtaining life time estimates of the bar by combining the effects of crack initiation and propagation. A description of the S/N Curve is shown in figure 6, where the base line data was obtained from a literature survey for material metallurgically similar to the torsion bar material. The survey provided a set of S/N Curves for torsional fatigue shown below for best representing the current materials used in the bar.

$$\log_{10} N = B + .068 \Delta\theta \quad (7)$$

where $B = 7.70$

The slope value of .068 was essentially the same for all curves in the set. The adjustment in B from 7.70 to 8.06 made on the basis of M60 torsion bar quality assurance tests at a single $\Delta\theta$ value performed at the Scranton manufacturing facility (See figure 6). A single load equivalent to a 42 degree angular displacement was applied during the quality assurance torsional fatigue test. Using the mean value and the cycles to failure in Figure 7 provided a more accurate estimate of (B). The curves representing a range of 10 and 20 percent reduction in bar stress are shown in figure 6.

The S/N Curve Monte Carlo process is similar to the previously outlined method for da/dN relationships. The primary difference involves using Models for (B) and $\Delta\theta$ from figure 6 and 2 respectively. A schematic of the basic S/N representation is shown in figure 8a and 8b. In figure 8a simulation of S/N curve variability is shown for a specific value. Figure 8b describes probability density function (PDF) for (B). A random selection of a discrete set of numbers from $\Delta\theta$ and (B) distributions is then applied to equation 7 in order to obtain $\log_{10} N$ value. The process is repeated until all values from the two distributions are selected. This process will then provide a PDF to represent $\log_{10} N$.

Torsion Bar System Reliability

By assuming a tank with a N torsion bar system the following procedures would be applied in order to establish reliability of the system. If any one bar could cause failure (independence) then reliability R will be

$$R = \prod_{j=1}^N P_j \quad \left\{ \begin{array}{l} P_j - \text{Prob. of Survival} \\ j - j^{\text{th}} \text{ Torsion Bar} \end{array} \right. \quad (8)$$

if it is assumed that all torsion bars must fail for system failure (dependence) then,

$$R = P_1 \times P_2 / P_1 \times \dots \times P_N / P_{N-1} / \dots / P_1 \quad (9)$$

where $P_N / P_{N-1} / \dots / P_1$ is the reliability of Nth bar, given reliabilities of N-1 ---, 1 bars.

Reliability of Operation After Specific Number of Cycles

The reliability of operating an additional number of cycles when a specified number of cycles of operation has been completed is obtained in the following manner. Initially it is assumed that a specified distribution function say $f(N)$ is known. For example the distribution of $\log_{10} N$ from Monte Carlo method previously described. The reliability $R(n_1, n)$ is a conditional probability requiring the probability of operating for $n_1 + n$ cycles when n_1 cycles have been completed. That is

$$R(n_1 + n) = \frac{R(n_1 + n)}{R(n_1)} = \frac{\int_{n_1 + n}^{\infty} f(N) dN}{\int_{n_1}^{\infty} f(N) dN} \quad (10)$$

where n is the additional mission in cycles after n_1 , cycles of operation. The number $N_s(n_1, n)$ of components (torsion bars) that will survive an additional n cycles is given by

$$N_s(n_1, n) = N_s(n_1) \cdot R(n_1, n) \quad (11)$$

where $N_s(n_1)$ = number of components starting the mission of n additional cycles.

Results and Discussion

The proper number of simulations for the Monte Carlo Method depended on the models under consideration. For example 5000 and 3000 were required for the da/dN and S/N curve models respectively. Using a convergence rate criteria for the calculated 1 percent values (see P_s in figure 9) and recognition of the third and fourth moment stability of the $\log_{10} N$ distribution provided an excellent method for determining required number of simulations. Differences in percentile values for C.V.'s of 10 and 15 percent were minimum. The 10 percent value was used for all $\frac{da}{dN}$ calculations.

The torsion bar reliability results from the da/dN relationship as shown in figure 10. The current design results were obtained from equation 3, with $A_2 = 3.29$. They indicated relative limited lifetime range of 14 to 500 miles, with a probability of survival values of .99 and .01 respectively. An appropriate increase in C_f from equation 3 represents the 40% increase in K_{IC} value. This represents an improvement in materials capability with respect to acceptance of larger flaw sizes prior to failure. The slight improvement in the bars capability indicates that an improvement in material will not significantly improve bar performance. The 25 and 50 % reduction in K_I (stress intensity) in figure 9 is obtained from reducing A_2 in equation 3 by the respective percentages. These reductions represent improvements in the design of spline section of the bar as shown in figure 1. The K_{III} failure in the shaft represents situations where failure occurs in shaft rather than spline region.

The maximum life of 70 miles at 25mph achieved from 50 percent improvement in spline design with .99 probability of survivability indicates that there is a very limited life of the bar after crack initiation. Table 2 describes minimum life estimates (99 percent survivability) for the torsion bar with respect to various tank velocities and the design improvements. Tank travel at 5mph (lowest speed) with a 50% reduction in K_I value shows propagation life expectancy of only 341 miles at .99 P_s .

In figure 11, the frequency distribution obtained from S/N curve - Monte Carlo application is shown. The resultant exponential form is consistent with that expected from the S/N modelled in the analysis.

A graphical display of P_s vs miles to failure is shown in figure 12 for the 25 mph tank velocity. The life expectancy of the bar is somewhat greater than that obtained from the da/dN analysis. The minimum life estimates (.99 P_s) of 292 miles is 21 times greater than 14 miles determined from the da/dN results. This result indicates that most of bar life occurs prior to crack initiation. Therefore the torsion bar should be manufactured in such a manner that flaws are minimized. The current shot peening used in the manufacture of the bar indicates recognition of this fact by the manufacturer. The bar reliability estimate obtained after an assumed 741 miles of tank travel (see figure 12), was obtained from equation 10. The increase in P_s from .90 to .99 if the bar survives the initial 741 miles does not provide a sufficient gain to warrant re-using bars since the minimum increase in expected life is reduced very rapidly. The results from a 20 percent reduction in design stress of 865 miles for a P_s of .99, is a considerable improvement when comparing that of 292 miles using in the current design. In table 3 the results from velocity ranging from 5 mph to 25 mph in increments of 5 mph are shown with respect to current 10 and 20 percent improvements in design. Reducing velocity of tank operation obviously improves reliability of the torsion bar. In this report, the experimental data and reliability calculations refer to failure of the first bar.

Examination of current design mileage capability of the bar for 20 and 25 indicates a range from 276 to 292 miles. These results agree with the 262 miles minimum life obtained from Aberdeen Proving Ground (APG) test results (Report MT-5376 of bar failure from 3 mile test course), (see figure 13). This course and tank velocity were similar to those used in obtaining the spectrum load results. The excellent agreement between the predicted and actual life expectancy of the bar indicates the desirability of Monte Carlo Process for modelling variability of spectrum loads (design stress) and S/N curve (material capability) results.

Although excellent agreement has been obtained, the authors would have preferred representing the spectrum load consistent with an individual peak to peak angular displacement. The simplification applied using the negative peak as base and representing the displacement relative to this value was a good approximation to the available individual displacements. This approximation would provide a slightly conservative estimate in the reliability values. Using the ASTM recommended practice of representing lower 3 standard deviation band of the S/N curve as measure of material fatigue loading capability combined with maximum angular displacement (46 degrees) for 25 mph. The tank operation resulted in a minimum life estimate of 112 miles for the bar. Selecting this number as a design allowable could result in an overly conservative estimate. The chance that this maximum displacement could occur and the S/N curve was the actual lower band described above is extremely small.

A minimum life of 575 miles was obtained from using the maximum $\Delta\theta$ displacement value with original S/N curve where $B = 8.06$. This result is obviously wrong since the limited samples of 23 bar failures two of them failed at mileage less than 400 miles (See figure 13).

Conclusions

1. A methodology for obtaining reliability of the M60 tank torsion bar subjected to cyclic random loads has been developed where probability of survival is represented as function miles of tank travel.
2. The developed methodology could be applied to other structures with cyclic random loads.
3. The use of the method appears justified from recognition of the excellent agreement between predicted reliability estimates and those obtained from the actual bar life (miles to failure) experienced during the tank operation.
4. Determination of minimum bar life was 21 times greater from application of S/N curve model than that of the assumed da/dN model. This indicates most of the bar life exist prior to crack initiation.
5. Application of deterministic procedures, (use of lower 3 S.D bound for S/N curve (ASTM method) and mean S/N curve providing over and under design allowable estimates while Monte Carlo method outlined in the text values accurately described acceptable design values.

References

1. D.M. Neal and D.S. Mason, "Determination of Structural Reliability Using A Flaw Simulation Scheme", Army Materials and Mechanics Research Center TR81-53, 1981.
2. "Damage Tolerant Design Handbook", Metals and Ceramics Information Center, Battelle Columbus Laboratories, MCIC-HB-01, Page 8.2-D, 1975.
3. Ibid. Page 8.2-E.
4. J.M. Barsom, Transactions of the ASME, Journal of Engineers for Industry, Series B, 93, No. 4, Nov. 1971.

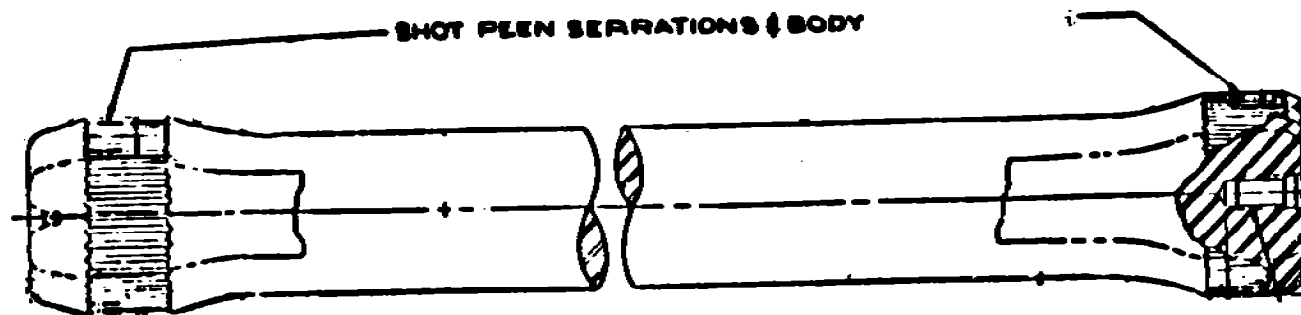


FIGURE 1 M60 TANK TORSION BAR

Load Spectrum (Run #48 - Speed 25 mph)

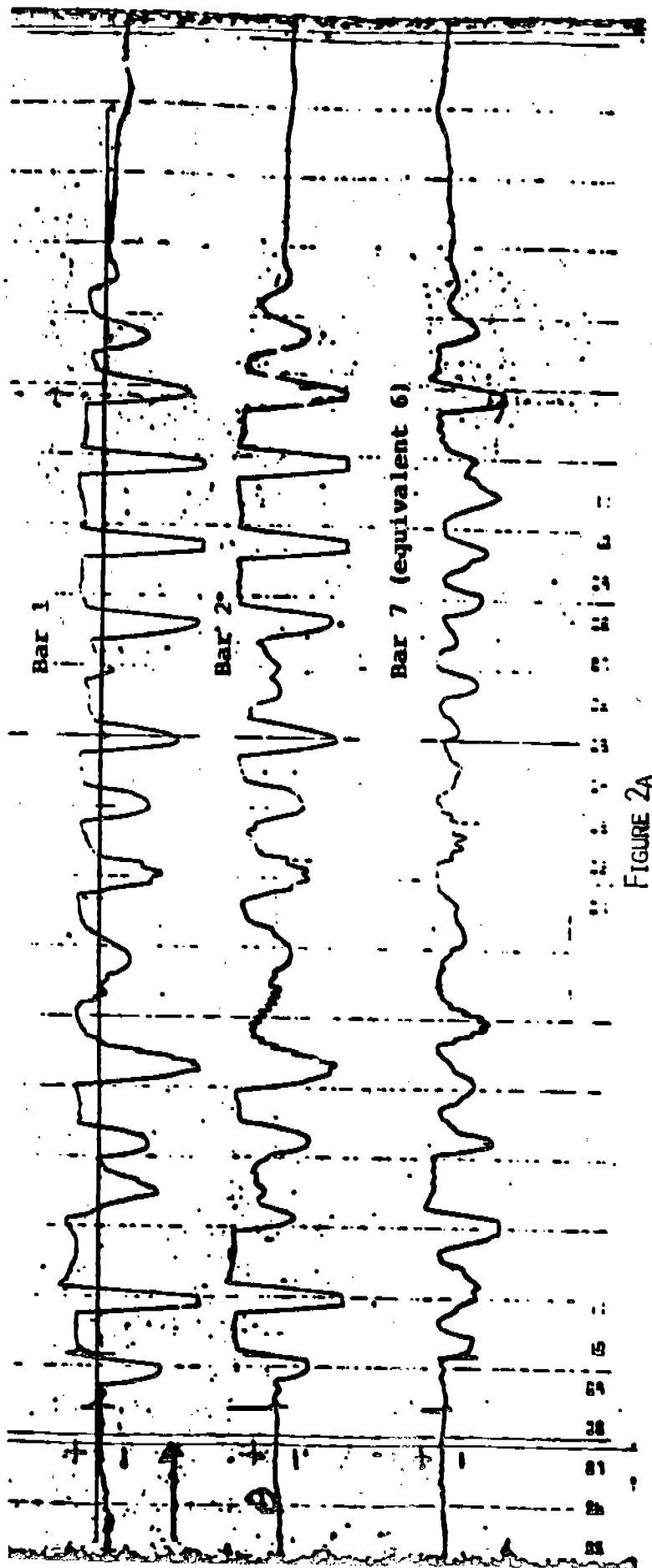


FIGURE 2A

Course Length 462 ft

AMPLITUDE DISTRIBUTION DATA RUN 48 (25 MPH)

	+ Peak	- Peak	+ 99%	- 99%	+ 66%	- 66%	+ 50%	+ 25%	- 25%
BAR 1	15.47	-30.71	15.33	-28.44	8.81	-3.51	3.00	-8.04	10.91
BAR 2	21.70	-28.92	21.42	-26.80	8.41	-7.43	-1.06	-11.53	13.92
BAR 7	13.91	-30.61	13.76	-30.47	6.63	-2.39	2.56	-7.19	8.35

FIGURE 2B

Spectrum Load Representation (Run #48 - 25 mph)

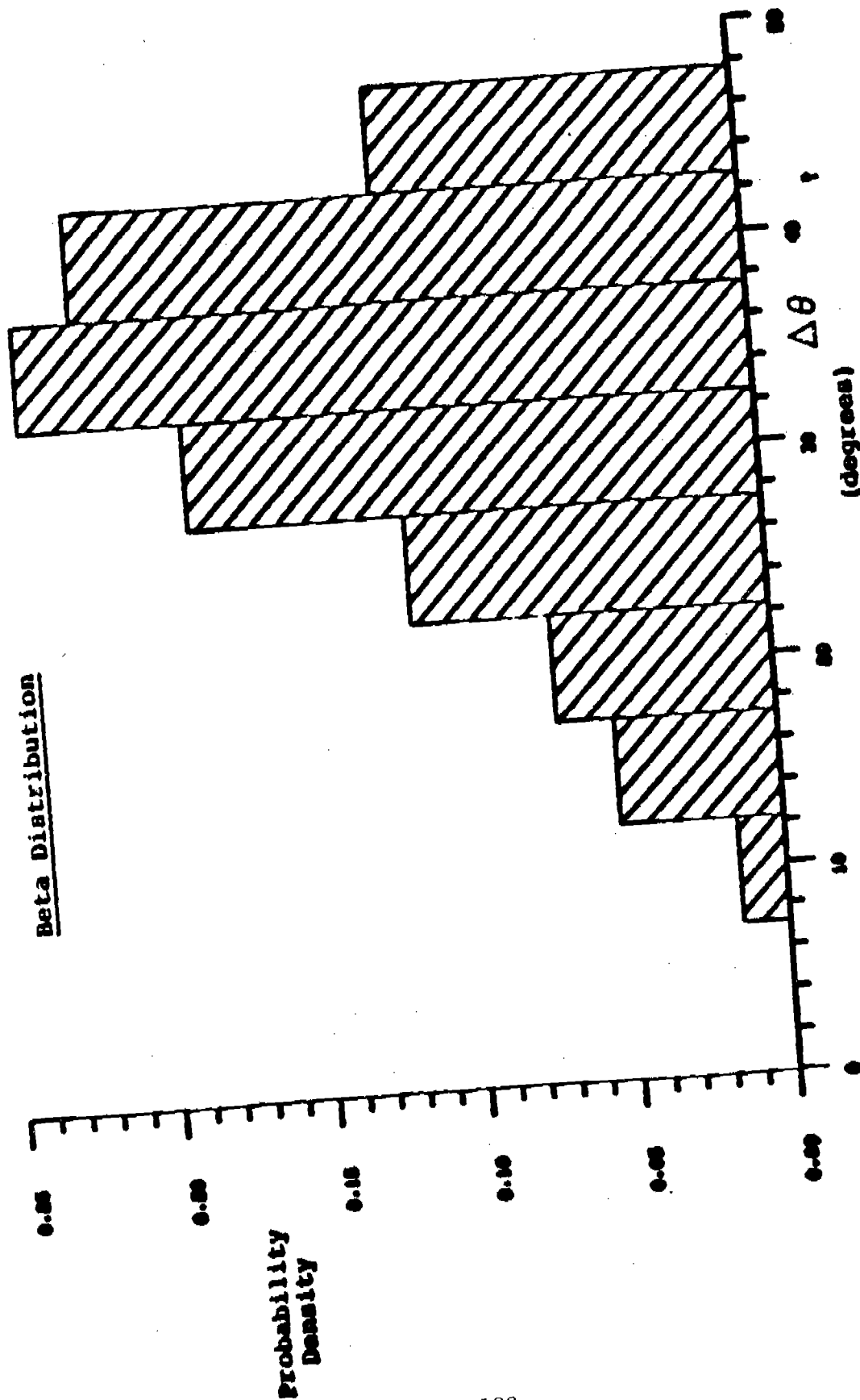


FIGURE 3

Determination of Cycles to Failure (LEFM Appr:)

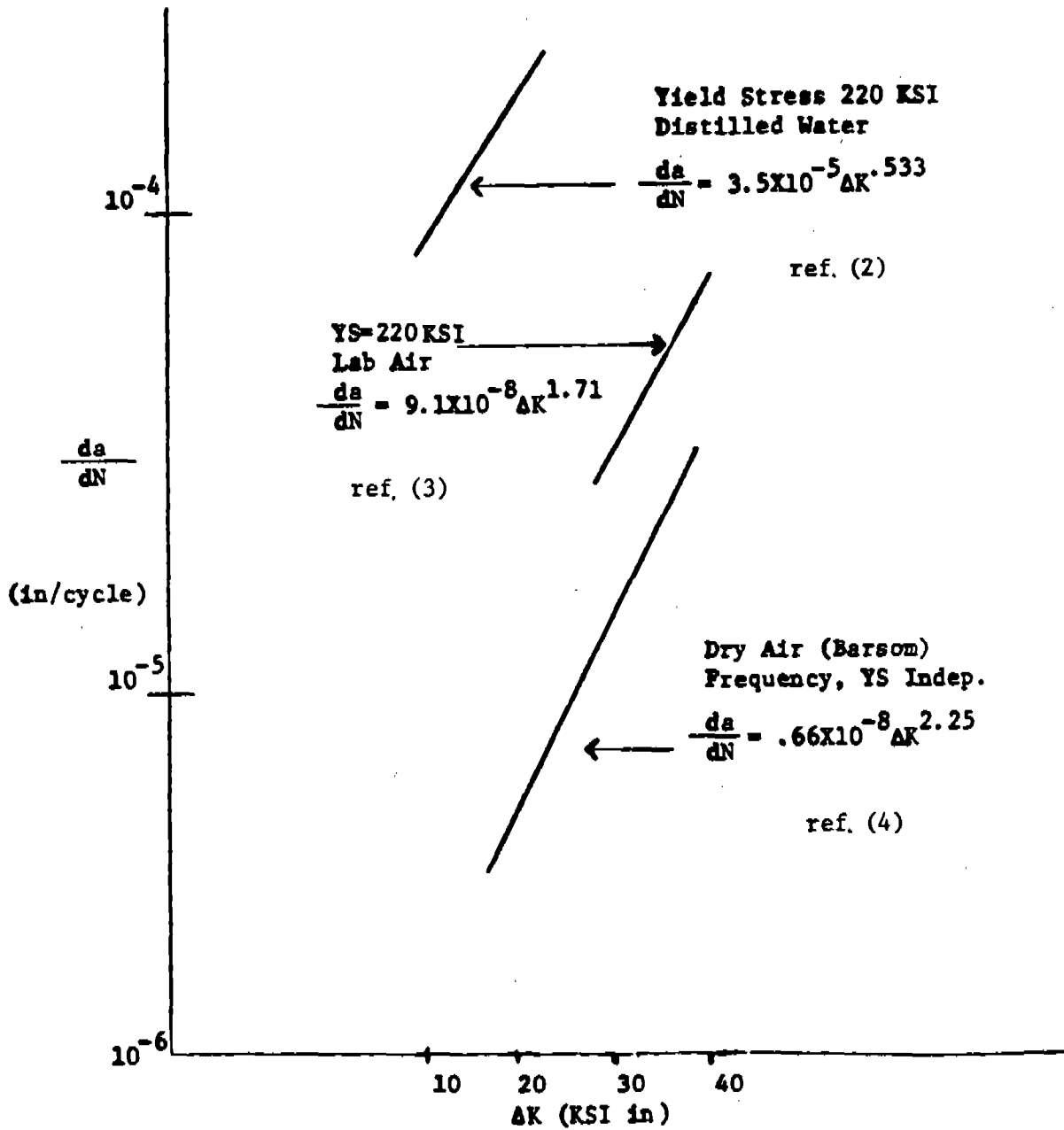
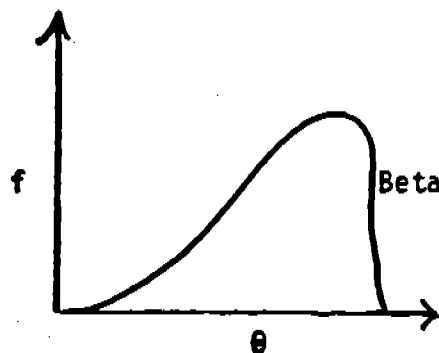
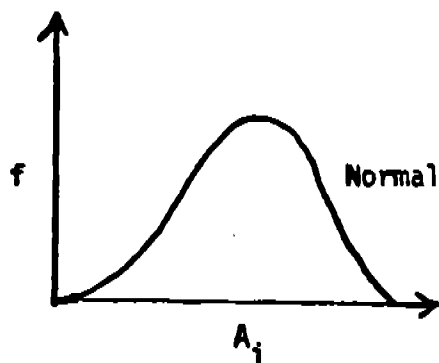
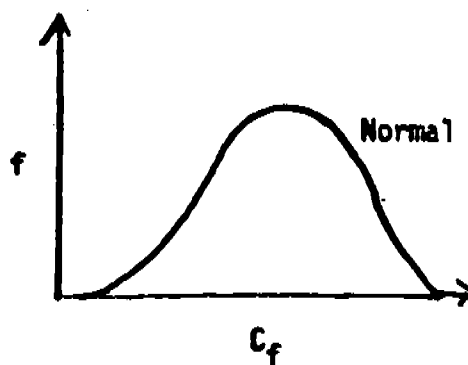
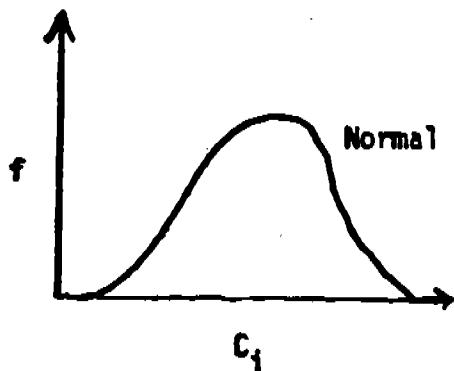


FIGURE 4



$$\int_{-\infty}^X f_i dX = R$$

R = UNIFORM RANDOM NUMBERS

f_i = FREQUENCY DISTRIBUTION

$$\bar{C}_i = .001 \text{ in.}$$

$$\bar{C}_f = .0133 \text{ in.}$$

$$\bar{A}_2 = 3.29 \text{ in. (Spline Region)}$$

[C.V. = 5, 10, 15 Percent]

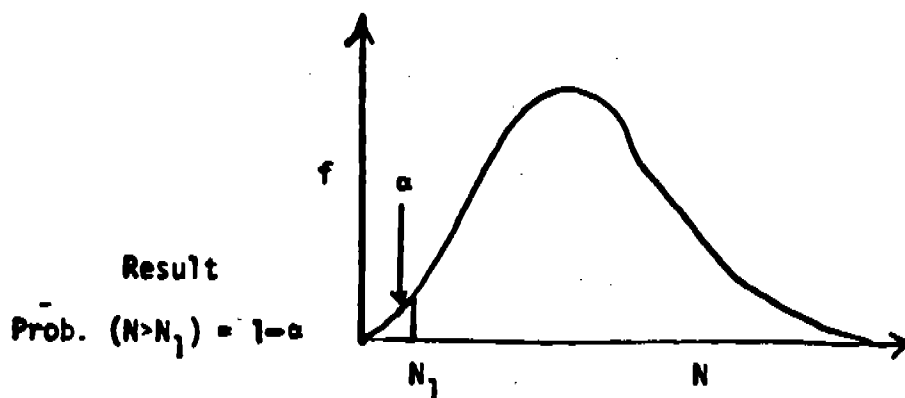


FIGURE 5
190

Torsional Fatigue Life

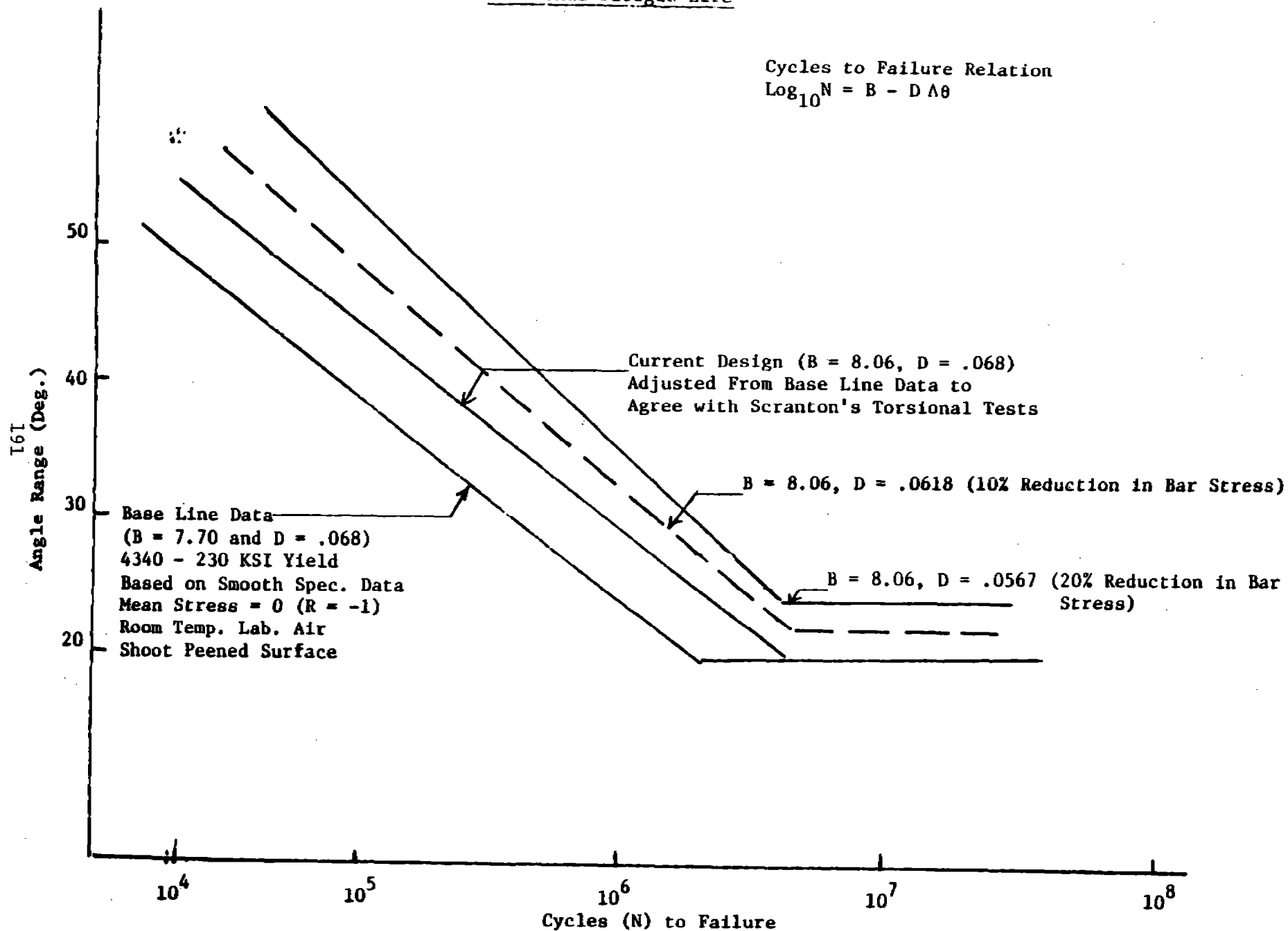
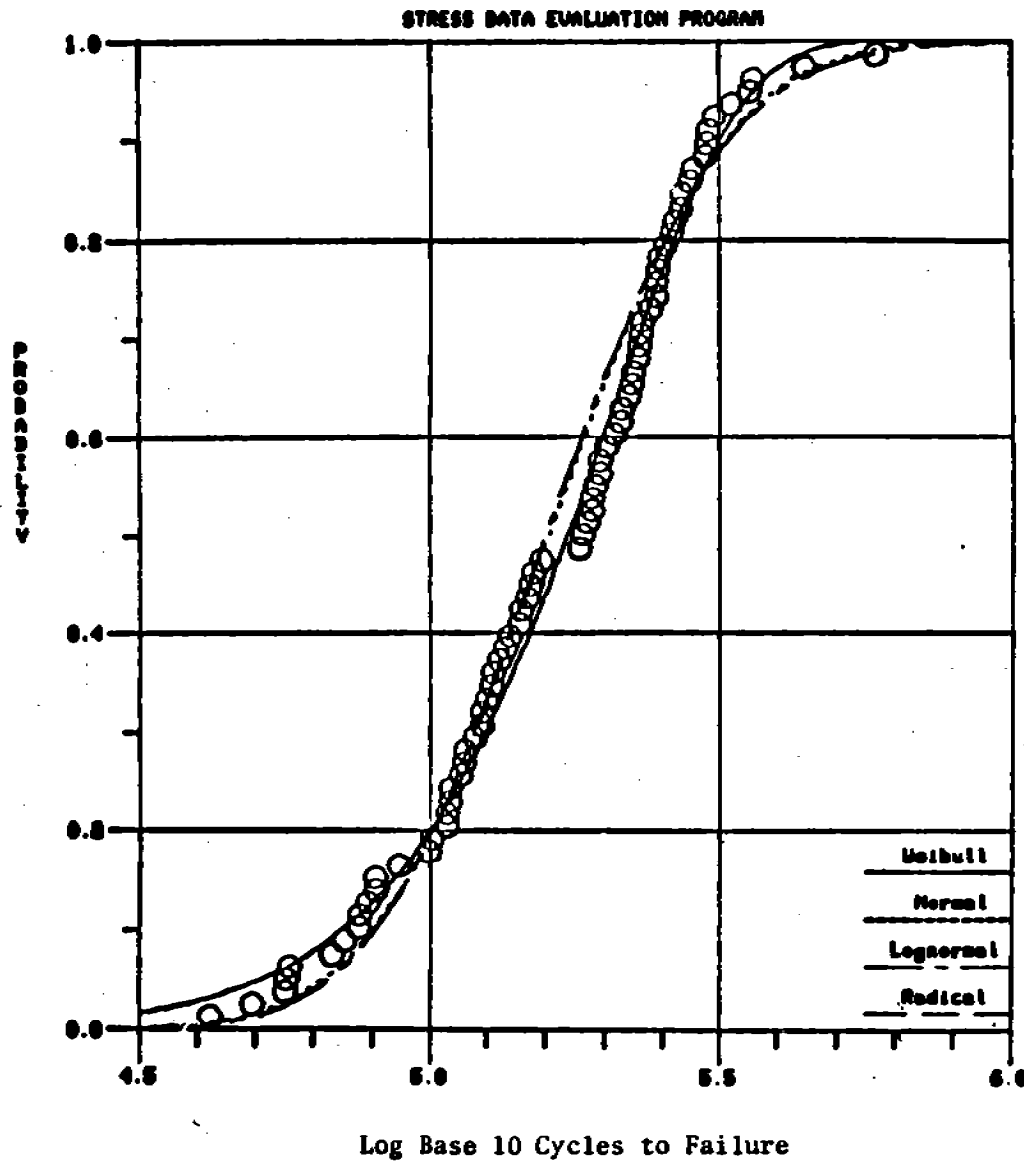


FIGURE 6



Mean = 5.206
Standard Deviation = .239

Design Allowable

A = 4.55
B = 4.83

FIGURE 7

Torsional Fatigue Life

Cycles to Failure Relation

$$\log_{10} N = B + .068 \Delta\theta$$

$B = 8.06$

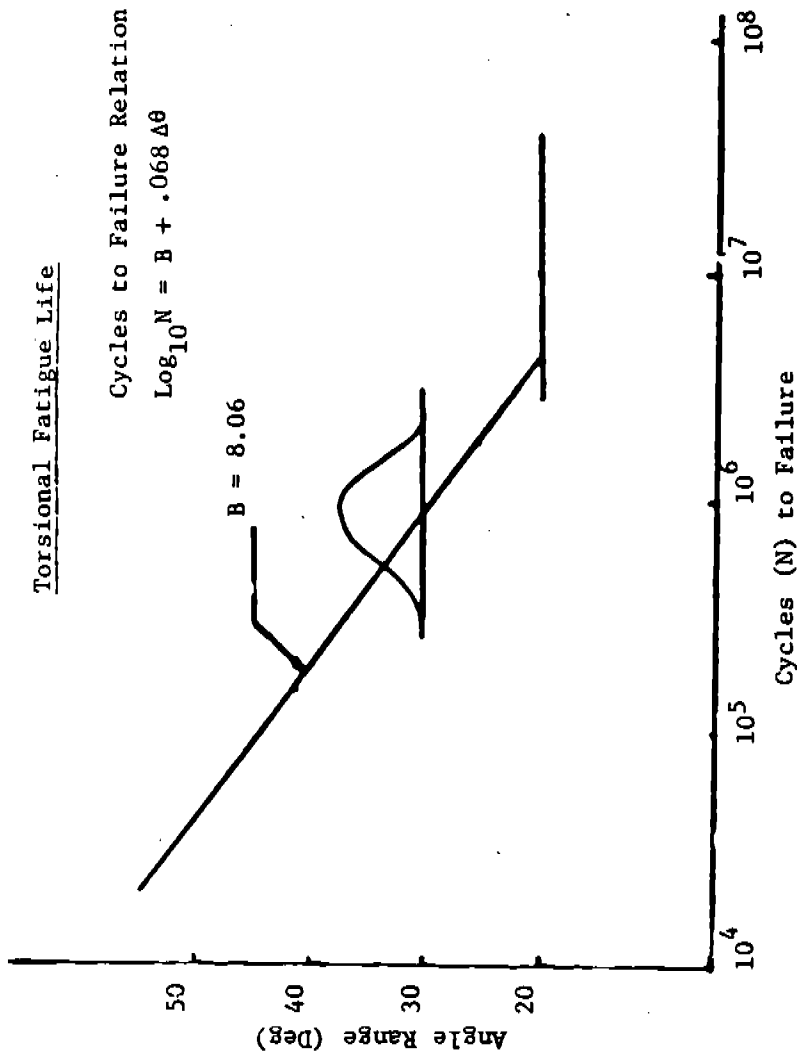
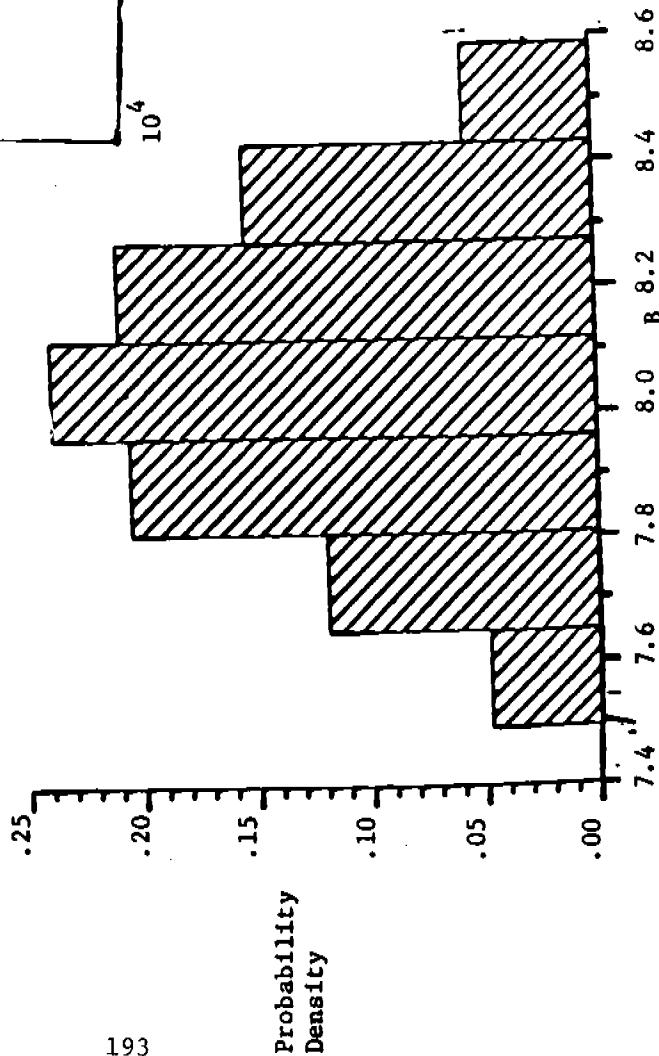


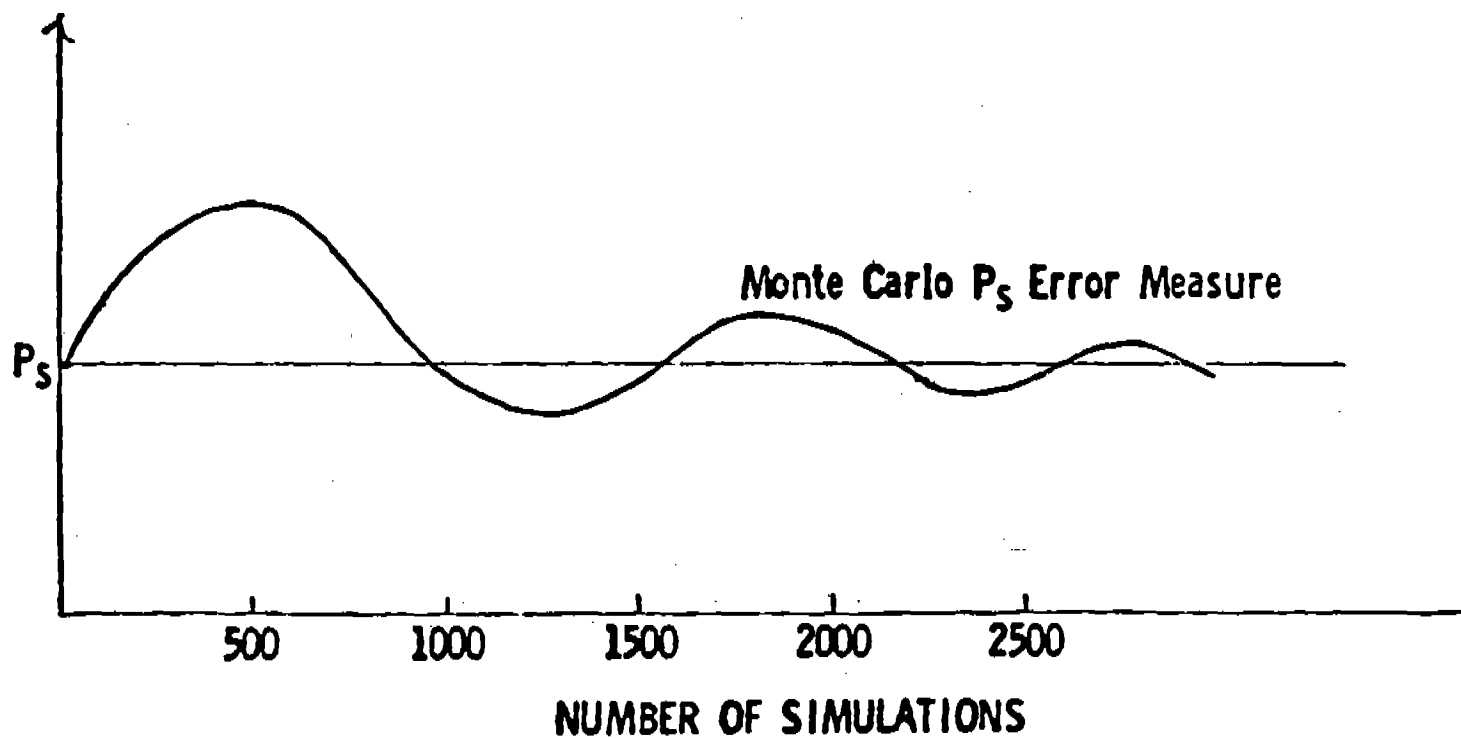
FIGURE 8A

Simulation of B from S/N Curve Representation



S/N Curve $\log_{10} N = B + D\Delta\theta$

FIGURE 8B



Additional Criteria: Convergence of 3rd and 4th Moments

FIGURE 9

Torsion Bar Reliability - Probability of Survival vs Miles da/dN Relationship

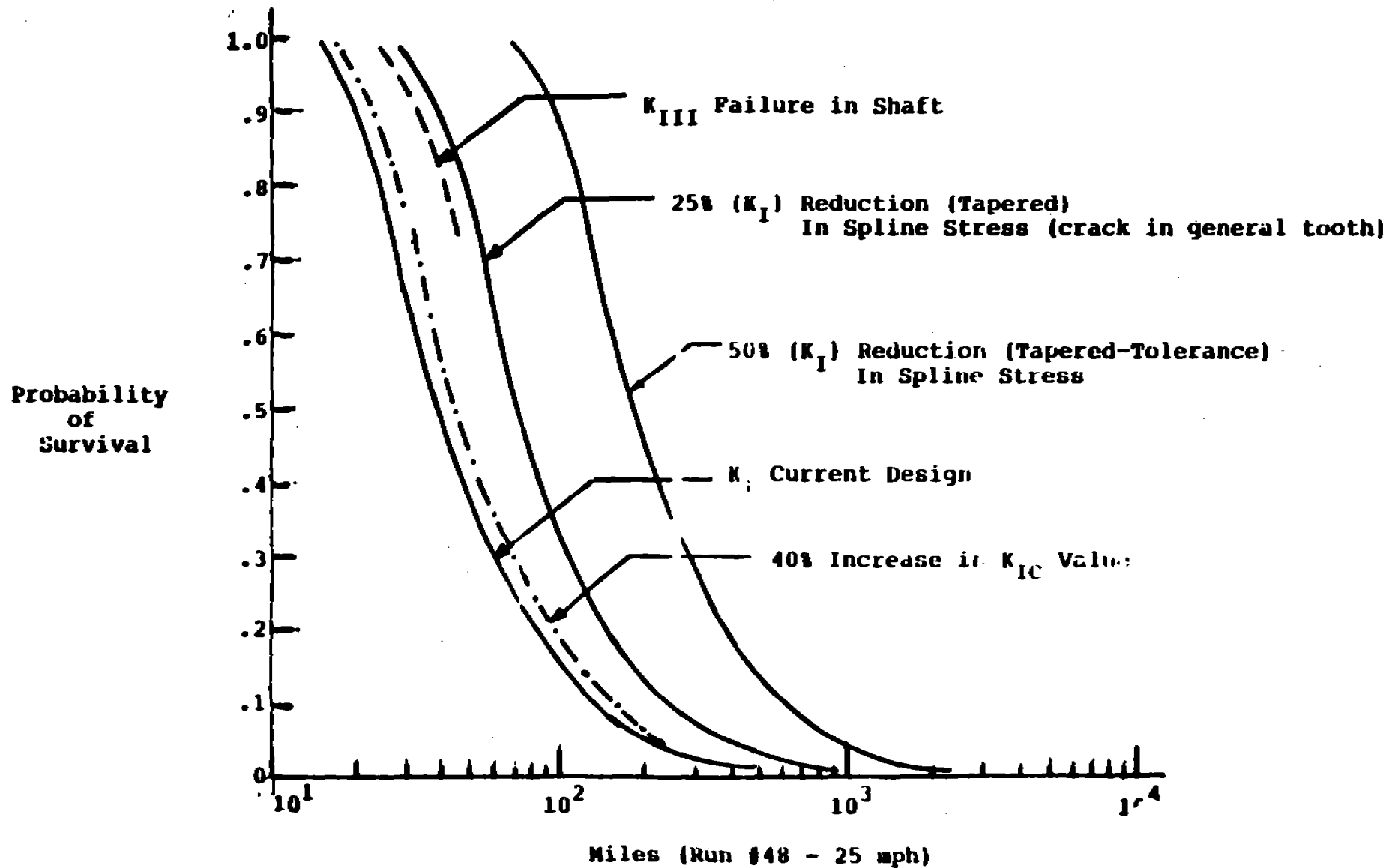
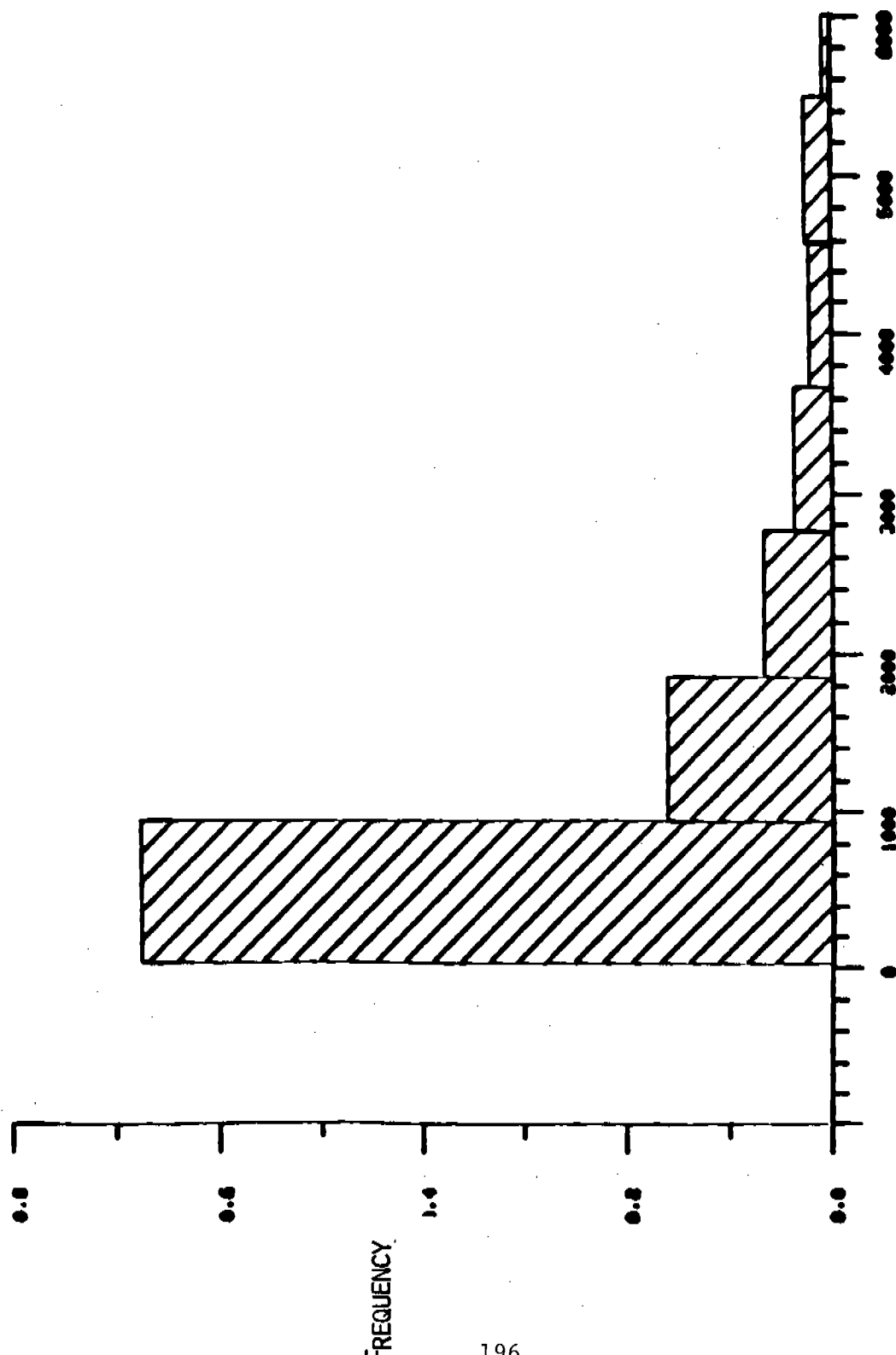


FIGURE 10



Miles to Failure #1 Torsion Bar (miles at 25 mph)

FIGURE 11

Probability of Survival vs Miles S/N Curve Results

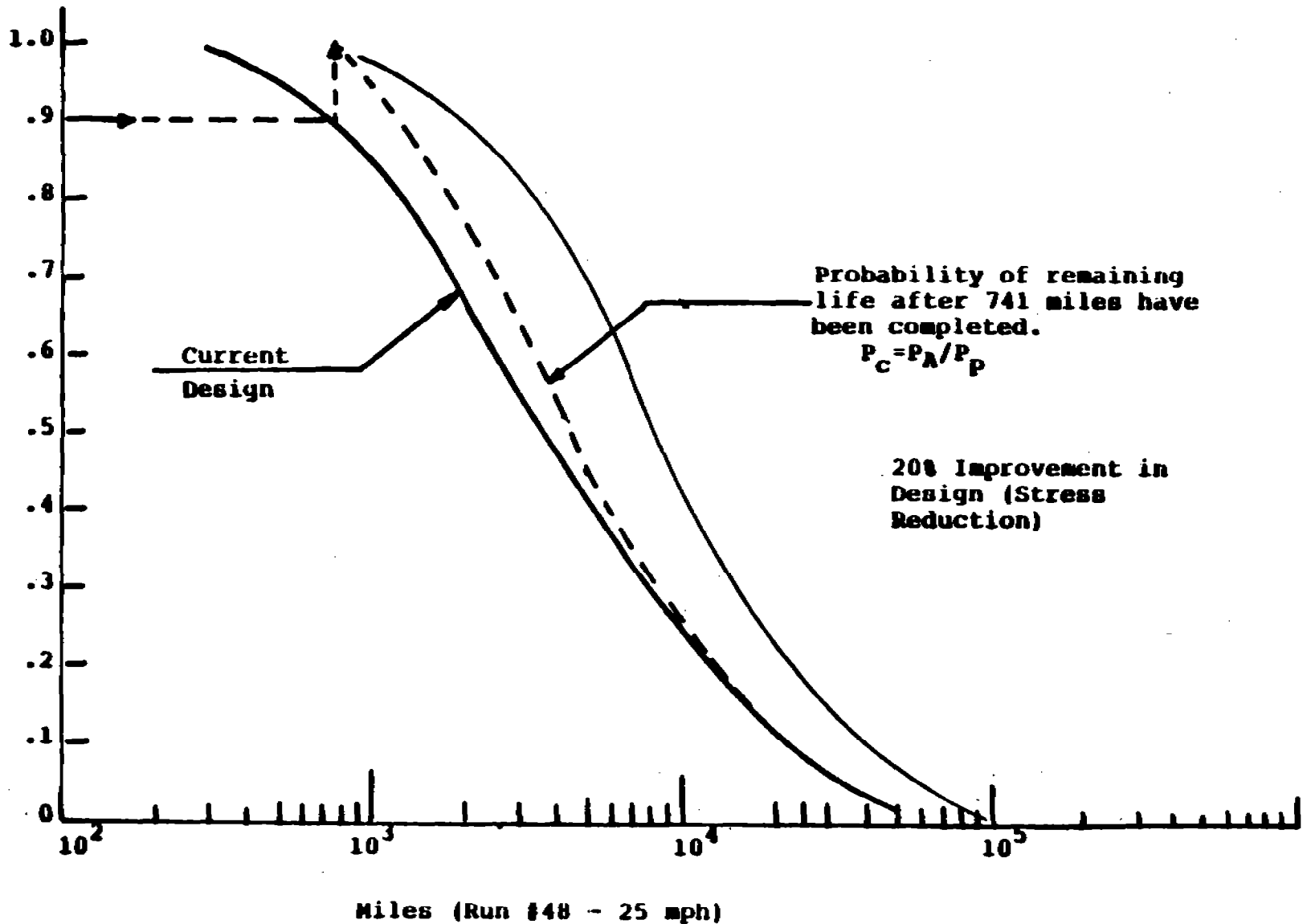
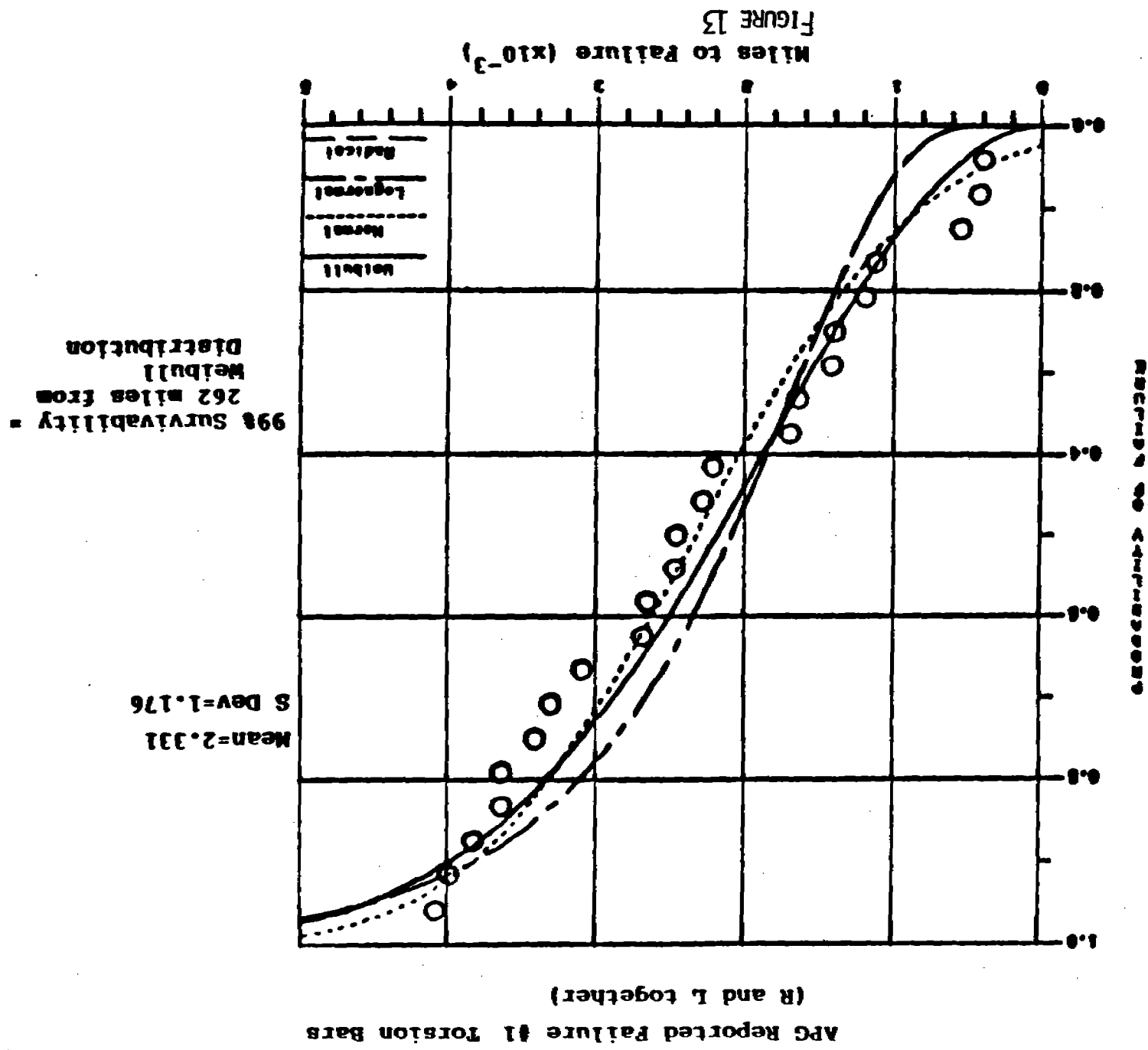


FIGURE 12



Spectrum Load (Profile IV Course) - Beta Function Representation

Cummulative Probability	θ (Degrees) Test Results	θ (Degrees) Beta Representation
Run 40 (5 mph)		
.10	.14	.86
.25	4.4	5.8
.34	5.0	7.2
.50	8.6	9.3
.66	12.5	11.5
.75	14.2	12.7
.99	17.0	16.7
Run 42 (10 mph)		
.10	14.0	6.1
.25	16.0	19.2
.34	22.8	21.5
.50	25.8	24.8
.66	29.7	27.6
.75	30.6	29.0
.99	32.6	32.5
Run 48 (25 mph)		
.10	2.3	10.8
.25	22.7	26.2
.34	27.2	29.1
.50	33.7	33.3
.66	39.5	37.1
.75	41.6	39.3
.99	46.0	46.9

Cummulative Time Probabilities of Torsional Bar Angular Displacement θ adjusted to positive range by $\theta = \theta + |\theta^-|$ where $\theta^- = \text{max. negative angular displacement.}$

TABLE 1

**Minimum Life Estimates (99% Survivability)
da/dN Curve Results**

Velocity (MPH)	Mileage Expected (Function of Spline Stress)		
	Current	25% Reduction	50% Reduction
5	71.0	138	341
10	29.9	51.9	143
15	15.2	29.3	72.3
20	14.0	26.9	66.3
25	14.2	28.4	70.2

TABLE 2

**Monte Carlo Results for S/N Curve Minimum Life
Estimate (99% Probability of Survival) vs Velocity (MPH)**

Velocity	Mileage	Expected	
(MPH)	Current Design	10% Design Improvement	20% Design Improvement
5	6,974	9474	12970
10	2,000	3138	4420
15	345	638	1089
20	276	515	860
25	292	557	865

***Note:** A 99% survivability estimate of 262 miles was obtained from cumulative APG mileage on vehicles at time of torsion bar failure. Velocity of vehicle during tests was approximately 15 to 25 mph.

TABLE 3

RANDOM NUMBERS FROM SMALL CALCULATORS

Donald W. Rankin
Army Materiel Test and Evaluation Directorate
US Army White Sands Missile Range
White Sands Missile Range, New Mexico 88002

ABSTRACT. Random number generators are notoriously wasteful of digits; however, applying an augmented precision technique to a linear congruential generator enables one to produce on even a small calculator a set of pseudo-random numbers which contains a useful number of elements. This paper sets forth such a method.

1. **INTRODUCTION.** Most modern computers and many programmable calculators include in their softwares a function for generating "random" numbers. Such numbers are required any time a "Monte Carlo" test technique is employed.

It is usual to tailor each algorithm to a specific type of use, and to a specific size of computer. Probably it is not feasible to transfer such a tailored algorithm to a calculator of smaller size--particularly to one of shorter word length.

Perhaps the most efficient and certainly the most popular of these algorithms is the "Linear Congruential Generator." Mathematically stated,

$$x_{i+1} \equiv (ax_i + c) \bmod m.$$

All quantities are considered to be integers. If the modulus m be taken as some power of ten (or of two if in binary), the modular operation is effected by simple truncation.

Most calculators have the ability to truncate at the decimal point. A decimal point, therefore, is inserted solely for this purpose. Conceptually, the numbers remain integers.

Given that the modulus m is some (positive integer) power of ten, it is found that the algorithm generates a full set of m integers (ranging from zero to $m-1$, inclusive) whenever both

$$\begin{aligned} a &\equiv 1 \pmod{20} \text{ and} \\ c &\equiv (1, 3, 7, \text{ or } 9) \pmod{10}. \end{aligned}$$

The selection of values for a and c is an important part of adapting the algorithm to a specific case.

2. PSEUDORANDOM NUMBERS. Let us suppose that we have defined a set of m integers, all different. A random selection from the elements of this set requires that for any element, the probability of selection be $1/m$. Since this probability remains unchanged for subsequent selections, sampling with replacement is indicated.

We wish to develop an algorithm that does not depend upon an outside stimulus. However, it remains necessary to provide a value for x_0 , so that the process can begin. This value should be an element of the set, but the choice can be arbitrary. It is called the "seed." After each x_i is computed and used, it serves in turn as the "seed" for the next calculation. To avoid repetition, some programmers employ a date-time group from which to extract a value for x_0 .

If any computed value of

$$x_i + s + 1 \equiv (ax_i + s + c) \pmod{m}$$

is ever equal to some previously used value of x_i , the algorithm will repeat itself over a subset of size $(s+1)$, exactly duplicating the previous cycle. If $x_{i+1} = x_i$, it is found that $s = 0$, and the algorithm has already degenerated into uselessness. To circumvent this, sampling without replacement is used. But this causes the probability of selection to increase as

$$\frac{1}{s}, \frac{1}{s-1}, \frac{1}{s-2}, \dots, \frac{1}{s-(s-1)}.$$

Thus, the last remaining element of the subset can be predicted with certainty. It is, of course, equal to x_i , the seed which began the cycle.

How then can we presume to use these sequences of numbers as "random" sequences? It is found that if the cycle length is very large (two hundred-fold would not be excessive) when compared with the quantity (of numbers) required, the sequence selected will exhibit certain of the characteristics associated with random sequences.

The term "pseudorandom" is used to indicate that the sequence is generated by an algorithm so that each element is a function of its predecessor.

3. PARAMETER SELECTION. At this point, let us limit the discussion to the case

$$m = 10^{2e},$$

"e" being a small, positive integer. Immediately

$$\sqrt{m} = 10^e.$$

It was observed in Section 1 that, under these conditions, maximum cycle length is achieved if c and m are relatively prime, and additionally $a \equiv 1 \pmod{20}$.

There are other requirements, however. Foremost among these is the restriction that ax_i must never overflow the computer word length. Should this occur, digits will be lost from the right, interrupting the flow of the algorithm and seriously shortening the cycle length.

The formula for serial correlation is

$$\rho = \frac{1 - 6 \left(\frac{c}{m} \right) \left(1 - \frac{c}{m} \right)}{a} + \epsilon$$

where $|\varepsilon| < \frac{a}{m}$.

It can be seen that the numerator varies from -0.5 to +1.0 and that the two terms are of the same order of magnitude when

$$a^2 \approx m.$$

The numerator can be reduced to zero by solving the associated quadratic in c/m . It is found that

$$\frac{c}{m} = \frac{1}{2} \pm \frac{1}{6} \sqrt{3}.$$

Now $\frac{1}{6} \sqrt{3} = 0.28867\ 51345\ 94812\ 88225\ 45743\ 90\ \dots$ is irrational, so that no element of the set can furnish a value for c which will reduce the numerator exactly to zero. It can, however, be made quite small, whence "a" can be set to a value somewhat less than $\sqrt{m}$ without adversely affecting the serial correlation.

At this point, it will be instructive to examine the sequence generated by the following parameters:*

$$\begin{aligned}x_0 &= 0 \\a &= 81 \\c &= 788677 \\m &= 1000000\end{aligned}$$

This sequence is found in Table 1-1. The entries are to be read as integers.

It is easy to observe that the least significant digit (units digit) is not "random" at all, since it can be predicted exactly. In the case at hand,

*All examples in this paper will assume an 8-digit calculator.

TABLE 1-1.
CYCLES OF DIGITS

a = 81

c = 788677

m = 1,000,000

$x_0 = 0$

0.788677	0.634957	0.137237
0.671514	0.220194	0.904874
0.181311	0.624391	0.083471
0.474868	0.364348	0.549828
0.252985	0.300865	0.324745
0.280462	0.158742	0.093022
0.506099	0.646779	0.323459
0.782696	0.177776	0.988856
0.187053	0.188533	0.886013
0.939970	0.059850	0.555730
0.926247	0.636527	0.802807
0.814684	0.347364	0.816044
0.778081	0.925161	0.888241
0.813238	0.726716	0.736198
0.660955	0.652835	0.420715
0.326032	0.668312	0.866592
0.197269	0.921949	0.982629
0.767466	0.466546	0.381626
0.953423	0.578903	0.700383
0.015940	0.679820	0.519700
0.079817	0.854097	0.884377
0.253854	0.970534	0.423214
0.350851	0.401931	0.069011
0.207608	0.345088	0.378568
0.604925	0.740805	0.452685
0.787602	0.793882	0.456162
0.584439	0.093119	0.737799
0.128236	0.331316	0.550396
0.175793	0.625273	0.370753
0.027910	0.435790	0.819670
0.049387	0.087667	0.181947
0.789024	0.889704	0.526384
0.699621	0.854701	0.425781
0.457978	0.019458	0.276938
0.884895	0.364775	0.220655
0.465172	0.335452	0.661732
0.467609	0.960289	0.388969
0.665006	0.572086	0.295166
0.654163	0.127643	0.697123
0.775880	0.127760	0.255640

it cycles through all ten digits, then repeats itself exactly. The two least significant digits, viewed as a single number, exhibit a similar cycle. A good generator will continue this effect until a cycle of length m is achieved. If m is a power of 10, this maximum cycle length is obtained whenever both of the following conditions are met:

1. $a \equiv 1 \pmod{20}$
2. c and m are relatively prime. This requires only that the final (units) digit of c be 1, 3, 7, or 9.

As an aid to continuing the study of the cycling effect, let us define

$$a_s = a^s \pmod{m}$$

and

$$c_s = c (1 + a + a^2 + \dots + a^{s-1}) \pmod{m}.$$

Given	$a = 81, s = 10, c = 788677$	we find
	$a_{10} = 928801$	
	$c_{10} = 939970$	

Note that, since $x_0 = 0$, c_{10} appears in the tenth position in Table 1-1. Now c_{10} may be viewed as having only five digits. It is therefore completely exercised by a five-digit multiplier, and we need merely use the last five digits of a_{10} . The parameters

$$\begin{aligned} x_0 &= 0 \\ a_{10} &= 28801 \\ c_{10} &= 93997 \\ m_{10} &= 100\,000 \end{aligned}$$

will generate the sequence $x_{10}, x_{20}, x_{30}, \dots$

At first glance, this appears to exceed the calculator word length. However, if we multiply $x_i (a_{10}-1)$ and then truncate, the algorithm will run without difficulty. To express the complete formula

$$x_{i+10} \equiv x_i (a_{10} - 1)(\text{mod } m) + x_i + c_{10} \pmod{m}$$

It is convenient to compute a_5 by means of the binomial expansion. Hence

$$\begin{aligned} (1 + 80)^{10} &= 1 + 10(80) + 45(6400) + \\ &+ 120(80)^3 + 210(80)^4 + \\ &+ 252(80)^5 + \text{immaterial terms} \end{aligned}$$

The previous strategem will thus be available whenever s is a multiple of ten. The sequence thus generated is found in Table 1-10.

In a similar manner, the procedure can be reiterated and the sequence $x_{100}, x_{200}, x_{300}, \dots$ generated. Required values of the parameters are:

$$\begin{aligned} x_0 &= 0 \\ a_{100} &= 8001 \\ c_{100} &= 5197 \\ m_{100} &= 10\,000. \end{aligned}$$

This sequence is illustrated in Table 1-100.

The process can be carried no farther. To do so results in $a_{1000} = 1$, and the algorithm degenerates to the successive multiples of x_{1000} . This can be observed by looking at every tenth entry in Table 1-100. The phenomenon can be called a "quasi-cycle" of length 1000 and additive constant 197. It appears that original values of "a" congruent to 1 (mod 100) will hasten this effect and therefore should be avoided. Further scrutiny reveals that the "quasi-cycle" is actually of length 500 and additive constant 598.5.

TABLE 1-10.
CYCLES OF DIGITS

$a_{10} = 28801$

$c_{10} = 93997$

$m = 100,000$

$x_0 = 0$

0.939970	0.058770	0.777570
0.015940	0.574740	0.733540
0.027910	0.026710	0.625510
0.775880	0.214680	0.253480
0.059850	0.938650	0.417450
0.679820	0.998020	0.917420
0.435790	0.194590	0.553390
0.127760	0.326560	0.125360
0.555730	0.194530	0.433330
0.519700	0.598500	0.277300
0.819670	0.338470	0.457270
0.255640	0.214440	0.773240
0.627610	0.026410	0.025210
0.735580	0.574380	0.013180
0.379550	0.658350	0.537150
0.359520	0.078320	0.397120
0.475490	0.634290	0.393090
0.527460	0.126260	0.325060
0.315430	0.354230	0.993030
0.639400	0.118200	0.197000
0.299370	0.218170	0.736970
0.095340	0.454140	0.412940
0.827310	0.626110	0.024910
0.295280	0.534080	0.372880
0.299250	0.978050	0.256850
0.639220	0.758020	0.476820
0.115190	0.673990	0.832790
0.527160	0.525960	0.124760
0.675130	0.113930	0.152730
0.359100	0.237900	0.716700
0.379070	0.697870	0.616670
0.535040	0.293840	0.652640
0.627010	0.825810	0.624610
0.454980	0.093780	0.332580
0.818950	0.897750	0.576550
0.518920	0.037720	0.156520
0.354890	0.313690	0.872490
0.126860	0.525660	0.524460
0.634830	0.473630	0.912430
0.678800	0.957600	0.836400

TABLE 1-100.
CYCLES OF DIGITS

$a_{100} = 8001$

$c_{100} = 5197$

$m_{100} = 10,000$

$x_0 = 0$

0.519700	0.307700	0.095700
0.639400	0.427400	0.215400
0.359100	0.147100	0.935100
0.678800	0.466800	0.254800
0.598500	0.386500	0.174500
0.118200	0.906200	0.694200
0.237900	0.025900	0.813900
0.957600	0.745600	0.533600
0.277300	0.065300	0.853300
0.197000	0.985000	0.773000
0.716700	0.504700	0.292700
0.836400	0.624400	0.412400
0.556100	0.344100	0.132100
0.875800	0.663800	0.451800
0.795500	0.583500	0.371500
0.315200	0.103200	0.891200
0.434900	0.222900	0.010900
0.154600	0.942600	0.730600
0.474300	0.262300	0.050300
0.394000	0.182000	0.970000
0.913700	0.701700	0.489700
0.033400	0.821400	0.609400
0.753100	0.541100	0.329100
0.072800	0.860800	0.648800
0.992500	0.780500	0.568500
0.512200	0.300200	0.088200
0.631900	0.419900	0.207900
0.351600	0.139600	0.927600
0.671300	0.459300	0.247300
0.591000	0.379000	0.167000
0.110700	0.898700	0.686700
0.230400	0.018400	0.806400
0.950100	0.738100	0.526100
0.269800	0.057800	0.845800
0.189500	0.977500	0.765500
0.709200	0.497200	0.285200
0.828900	0.616900	0.404900
0.548600	0.336600	0.124600
0.868300	0.656300	0.444300
0.788000	0.576000	0.364000

The conclusion to be drawn is this: Even though the values of a and c be chosen so that the algorithm generates the full cycle of m integers before repeating, the number of elements of any "useful" subset probably does not exceed $\frac{1}{2} \sqrt{m}$. What is needed is a device to increase the effective word length of the calculator. How this can be done forms the subject matter of the next section.

In summary, let us view the number $ax_i + c$ before truncation. Obviously, the left-hand (most significant) digits are lost via the modular operation, leaving

$$x_{i+1} \equiv (ax_i + c)(\text{mod } m).$$

Now the x_i can assume, at most, " m " different values. Therefore, since both " a " and " c " are fixed, the quantity $ax_i + c$ also can assume, at most, " m " different values. What this means is that, provided the values of " a " and " c " are selected to produce maximum cycle length, the act of truncation does not reduce the quantity of numbers--only their size. It also shuffles their order.

What remains is, of course, x_{i+1} . It is usual to regard several of the right-hand (least significant) digits as "not significantly random." They are retained, however, for smooth operation of the algorithm, and to ensure that the full complement of " m " different numbers is delivered.

4. AUGMENTED PRECISION ARITHMETIC.

Double precision arithmetic is available in the software of many computers, and even in some calculators. It is cumbersome to program and executes very slowly. This is particularly true with division.

However, the algorithm for the linear congruential generator does not employ division. Moreover, since $a^2 < m$, the word length ($m \sqrt{m} - 1$) is sufficient.

Let $m = 10^{2e}$, where e is any small positive integer. The "augmented" word consists of three parts, each of which consists of " e " digits.

Let us express x_i in the form

$$x_i = u_i \times 10^e + v_i$$

Thus a , u_i , and v_i are all integers less than $\sqrt{m}$, and the product of any two of them will not cause overflow.

Some calculators will compute (but not necessarily display) an extra digit. For them, the procedure is extremely easy. First, compute

$$(au_i \times 10^e)(\text{mod } m).$$

To this quantity, add $(av_i + c)$ and truncate again. The result is x_{i+1} .

When place for an extra digit is lacking, it is necessary to devise a procedure which avoids overflow. The following method, which assembles x_{i+1} by parts, beginning at the right, works quite well.

As before, express x_i in the form

$$x_i = u_i \times 10^e + v_i.$$

In analogous fashion, express " c " as

$$c = p \times 10^e + q,$$

Store p , q , u_i , and v_i separately. Select " a " so that

$$a < 10^e$$

$$a \equiv 1 \pmod{20}$$

$$a \not\equiv 1 \pmod{100}$$

It will be found that $a < 10^e - 18$. Consequently, multiplication by parts will not produce overflow.

We have immediately

$$v_{i+1} \equiv (av_i + q) \bmod 10^e.$$

Since we wish to retain both parts of $(av_i + q)$, we compute $(av_i + q) \times 10^{-e}$, then "FRC" $((av_i + q) \times 10^{-e})$.*

v_{i+1} is now stored, replacing v_i . Then $u_{i+1} \equiv (au_i + p + 10^e(av_i + q - v_{i+1})) \bmod 10^e$.

The sequence of numbers generated by

$$\begin{aligned}x_0 &= 0 \\a &= 9941 \\c &= 2113\ 2487 \\m &= 100\ 000\ 000\end{aligned}$$

is displayed in Table 2-1.

5. RANDOM SELECTION. RANDOM ORDERING.

So far, an algorithm has been developed which will generate a full set of m pseudorandom numbers. However, the length of a useful sequence of these numbers is, at best, uncertain and doubtless does not exceed $\frac{1}{2}\sqrt{m}$.

If a subset of far smaller but exactly known size is to be placed in random order, or if random selections from its elements are to be made, the following can be done.

*"FRC" means "fractional part of."

Storage space must be provided to accommodate all the elements of the subset, plus one more. It may be possible that scratch-pad storage is adequate.

Let us illustrate the method by example. Suppose the task at hand is to shuffle a pack of 52 playing cards, i.e., to place them in random order. We thus require 53 storage registers, which we number from 00 to 52, inclusive. The individual card names are entered into registers 00 through 51 in any arbitrary order. $N = 52$ is the subset size.

We employ the generated sequence of numbers given in Table 2-1. These numbers (integers) should be distributed uniformly on the interval 0 to m . Dividing by m , then multiplying by 52, yields a sequence uniformly distributed on the interval 0 to 51.99999 The "integer" portion of this number is used as an address for selecting a card. That card is then placed in storage register 52.

Next, all cards with location numbers greater than the "selected" location are cascaded downward one position. This includes the card placed in register 52. So far, the illustrative example has given $52 \times 0.21132487 = 10.988$ The card in location 10 was drawn and stored in location 52. Say it is the Spade Jack.

After cascading, only 51 cards are of interest. Hence $51 \times 0.99185754 = 50.584$ The card in location 50--the King of Clubs--is drawn and placed in register 52. Again after cascading, the subset of unshuffled cards is reduced to 50 in number. Hence $50 \times 0.26713001 = 13.356$ The card now in location 13--the deuce of Hearts--is drawn and placed in register 52.

Continuing as above, $49 \times 0.75075428 = 36.786$ The card in location 36--say the King of Diamonds--is selected and placed in location 52.

When the size of the unshuffled subset is reduced to unity, that card certainly will be found in location 00, and it certainly will be selected for transfer to location 52. Consequently, that transfer can be effected without

TABLE 2-1.
CYCLES OF DIGITS

a = 9941

c = 2113 2487

m = 100,000,000

$x_0 = 0$

0.21132487
0.99185754
0.26713001
0.75075428
0.45962235
0.31710622
0.56425789
0.49900936
0.86337263
0.99863970

0.24103567
0.34692034
0.94642401
0.62036108
0.22082115
0.39437702
0.71328069
0.93466416
0.70773943
0.84899850

0.22722647
0.06966314
0.73259961
0.98404788
0.63129995
0.96412782
0.60598349
0.29319896
0.90218623
0.84463730

0.68858257
0.41065324
0.51518371
0.65258598
0.56855205
0.18725392
0.70254359
0.19715306
0.10989433
0.67085940

0.10541337
0.12563604
0.15919851
0.80371278
0.92007085
0.63564472
0.15548639
0.90152786
0.29978113
0.33553820

0.75072417
0.16029884
0.74209331
0.36091958
0.11286965
0.24851552
0.70410919
0.76078266
0.15174793
0.73749700

0.22462027
0.16142894
0.97641741
0.77679768
0.35706175
0.76218162
0.05880929
0.83447676
0.74479603
0.22865910

0.79657107
0.92433174
0.99315221
0.13744448
0.54690055
0.94969242
0.10367209
0.81557156
0.80820283
0.55565790

0.66900187
0.75891454
0.58076701
0.61617128
0.57001935
0.77368322
0.39621489
0.98354636
0.64568963
0.01193670

0.31143797
0.21618464
0.30283111
0.65538938
0.43715145
0.93388932
0.00505499
0.46298046
0.70007773
0.68403880

0.00650877
0.91500744
0.30028591
0.35355618
0.91331025
0.42852012
0.12983779
0.92879526
0.36500453
0.72135760

0.87405957
0.23751024
0.30062071
0.68180298
0.01474905
0.83163092
0.45430059
0.41349006
0.71601133
0.07995640

employing the algorithm. Further, cascading can be omitted or not, at the pleasure of the programmer.

The final result is the shuffled deck, in order of selection, in the designated storage locations. Omitting the final cascading, the example leaves the Spade Jack in 01, the Club King in 02, the Heart deuce in 03, the Diamond King in 04, etc. The shuffled deck can now be put to the use for which it was intended.

If there is a requirement to "deal" the cards one at a time, it is suggested that the card in the highest numbered location be taken first. Not only is the programming simpler, but the stigma is avoided which usually is attached to dealing from the bottom.

In summary, a set of uncertain size has been used to produce a much smaller subset of known, fixed size.

6. STATISTICAL TESTS. There is much to be found in the literature on the subject of testing sequences of numbers to determine whether or not a sequence could have been produced by a random selection process. These methods will not be repeated here.

It is enough to be reminded that the answers to these statistical tests will be stated as probabilities. We should read nothing into the result beyond the probability statement itself.

BIBLIOGRAPHY

- (1) Abramowitz, M. and Stegun, I.A., eds., HANDBOOK OF MATHEMATICAL FUNCTIONS; Dover Publications, Inc., New York, 1972. (Sec 26.8)
- (2) Hull, J. E. and Dobell, A.R., RANDOM NUMBER GENERATORS; in SIAM Review, Vol 4 No 3 July 1962.
- (3) Jansson, B., RANDOM NUMBER GENERATORS; Almqvist and Wiksell, Stockholm, 1966.

APPLICATION OF THE BOOTSTRAP METHOD
TO A MEASURE OF FORCE EFFECTIVENESS
(AN EMPIRICAL CASE STUDY)

Eugene Dutoit
Ellen Shannahan
OR/SA Branch
Directorate of Combat Developments
US Army Infantry School
Fort Benning, Georgia 31905

and a

Postscript submitted by
Joseph Tessmer
1947th HQ Support Group
SAGF

Directorate for Theater Force Analysis
Fighter Division
USAF
Washington DC 20330

ABSTRACT. A paper was presented at the Twenty-Eighth Conference on the Design of Experiments about estimating the variance of the loss exchange ratio (LER). The LER is a measure of force effectiveness that is often used in military analysis of combat. Two methods of estimation were discussed: (1) the method of error propagation, and (2) the application of Fieller's theorem. The discussion that followed the presentation and further references to the literature pointed to Fieller's method as the preferred methodology to use to estimate confidence intervals about this measure of force effectiveness. Professor Bradley Efron (Stanford University) presented an overview of bootstrap methods. Dutoit and Shannahan have applied bootstrap methods to data to compute an estimate of the LER. Confidence intervals were also determined. The distribution of LERs about the mean value derived from the bootstrap have been compared to results using error propagation and Fieller's theorem. The results of this comparison as well as the bootstrap sensitivity to different replication sizes are presented.

1. INTRODUCTION AND BACKGROUND.

a. Error Propagation and Fieller. As pointed out in reference (2), the LER is defined as the ratio of Red casualties (R) to Blue casualties (B):

$$\text{LER} = R/B. \quad (1)$$

Usually the values of R and B are obtained by replicating a stochastic wargame model. The average LER ($\bar{\text{LER}}$) is computed as:

$$\bar{\text{LER}} = \bar{R}/\bar{B} \quad (2)$$

Because the generators of these average values are the results of a stochastic

wargame, it would be useful to determine a confidence interval around the measure for various forms of hypothesis testing. Using error propagation methods, reference (2) shows that the variance of the (LER) can be estimated as:

$$\text{VAR}(\hat{\text{LER}}) = \frac{1}{n} \left[\left(\frac{1}{\bar{B}} \right)^2 S_R^2 + \left(\frac{-\bar{R}}{\bar{B}^2} \right)^2 S_B^2 + 2 \left(\frac{1}{\bar{B}} \right) \left(\frac{-\bar{R}}{\bar{B}^2} \right) R S_R S_B \right] \quad (3)$$

The appropriate $100(1 - \alpha)$ confidence interval (C.I.) for the LER would be calculated as:

$$100(1 - \alpha) \text{ C.I. (LER) } = \hat{\text{LER}} \pm t \sqrt{\text{VAR}(\hat{\text{LER}})} \quad (4)$$

Similarly, reference (2) also shows that Fieller's theorem can be used to find the fiducial limits of the ratio of two means. In this case, the upper and lower limits ($R_{U,L}$) can be found as the solution of a quadratic equation and are:

$$R_{U,L} = \frac{\bar{B}\bar{R} - t^2 R S_B S_R}{n} \pm \frac{\sqrt{\left(\bar{B}\bar{R} - t^2 R S_B S_R \right)^2 - \left[\bar{B}^2 - t^2 \left(\frac{S_B^2}{n} \right) \right] \left[\bar{R}^2 - t^2 \left(\frac{S_R^2}{n} \right) \right]}}{\bar{B}^2 - t^2 \left(\frac{S_B^2}{n} \right)} \quad (5)$$

In operations (2), (3), (4), and (5) the following notation is used:

- (a) $\bar{R}$, $\bar{B}$ are the average number of Red and Blue casualties, respectively.
- (b) n is the number of stochastic wargame replications. This is used to calculate $\bar{R}$, $\bar{B}$, S_B , S_R , and R .
- (c) S_R , S_B are the sample standard deviations for Red and Blue casualties.
- (d) R is the correlation between Red and Blue casualties based on n replications of the wargame.
- (e) t is the two tailed value of the student's t with $(n-1)$ degrees of freedom.

The discussion that followed the presentation of this paper and further references to the literature pointed to Fieller's method as the preferred way (compared to error propagation) to compute a confidence interval about a ratio although there was an indication that both error propagation and Fieller's method to give "reasonably" consistent results.

b. Bootstrap. The purpose of this paper is not to provide a detailed description of bootstrap methods. Reference (1), entitled "Computer-Intensive Methods in Statistics" is a readily available and clearly worded explanation of the bootstrap method co-written by one of the bootstrap inventors (Efron). Figure 1 below shows how the bootstrap method was applied to sets of data to compute estimates of the LER and the frequency distribution of these estimates.

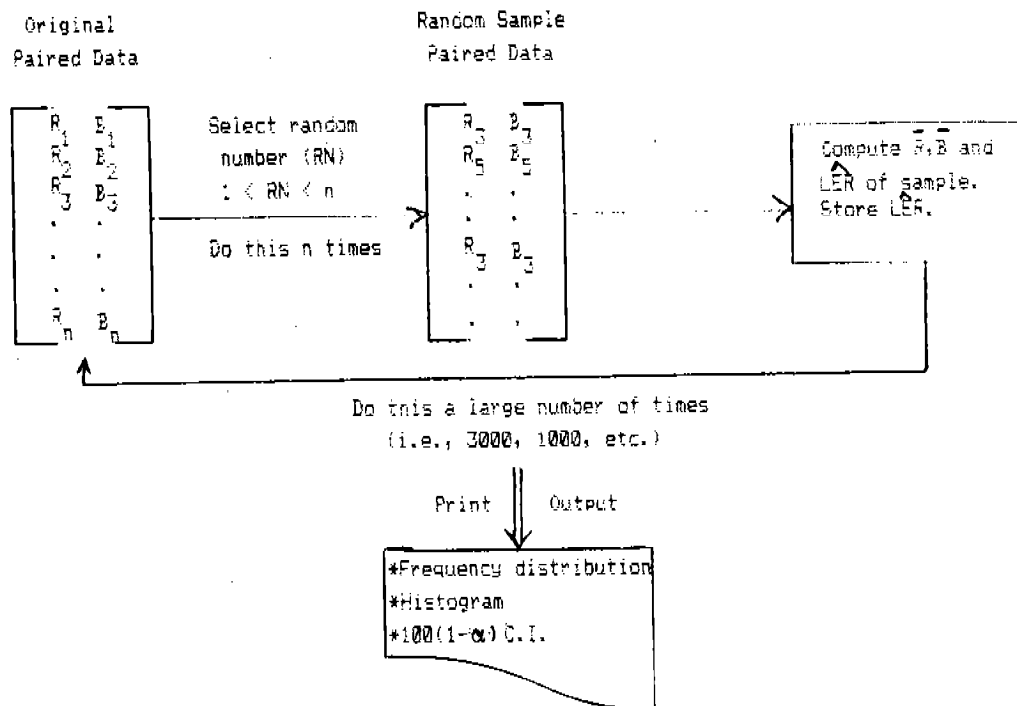


FIGURE 1. APPLICATION OF THE BOOTSTRAP METHOD TO ESTIMATE THE LOSS EXCHANGE RATIO (LER)

Each replication (1,2,3,...,n) of the stochastic wargame provides a set of paired data (i.e., Red and Blue casualties). Through random selection (with replacement), another set of data of n paired observations is selected from the original data set. From this additional sample, the values of $\bar{R}$ and $\bar{B}$ are obtained and the value $\hat{LER}$ is computed. This value of $\hat{LER}$ is stored in the computer memory. This bootstrap process is done a large number of times (3000, 1000, etc.) and the sample $\hat{LER}$ is stored for each additional sample. At the completion of a large number of bootstrap runs, the frequency distribution is printed and the average LER, as well as the appropriate confidence limits, are determined from this empirical distribution. These LER estimates, and the confidence limits derived from the bootstrap, were compared to results using error propagation and Fieller's theorem. The results of this comparison as well as the bootstrap sensitivity to different replication sizes (3000, 1000, 750, 500, 250, 100) was studied.

2. ASSUMPTIONS AND CONSTRAINTS. The following assumptions and constraints apply to this study.

a. This is a case study based on actual data obtained from the CARMONETTE stochastic wargame. The findings or observations should be interpreted as emerging trends with respect to LERs within the constraints of the forces and systems modeled using this wargame. Perhaps this paper will serve as a catalyst for some additional theoretical studies using bootstrap methods to estimate measures of force effectiveness.

b. It is assumed that the Radio Shack TRS-80 Model II (64K) system random number generator produces a statistically valid stream of 36,000 random numbers (minimum).

c. The emerging findings or trends apply to 99, 95, 90 and 85% confidence intervals.

d. Estimates of the average value of the $LER(\hat{LER})$ are carried out to the nearest tenth. This measure of force effectiveness is a rough indicator and estimates made with any greater precision are not considered to be operationally meaningful.

3. THE SOURCE OF DATA USED IN THIS STUDY. The data used in this case study were obtained from a force-on-force evaluation of several medium antiarmor systems which were employed within an Infantry force and scenario. Twelve medium antiarmor concepts were examined (denoted as case A, B, C, ..., L). All medium antiarmor concepts were inserted in the same force and fought against the same threat on the same terrain. All other factors were held constant, therefore the differences in average Red and Blue casualties are attributed to the performance factors and synergistic influence of the different antiarmor systems. Table 1 below shows the input to this case study for each of the twelve antiarmor systems (cases A through L). The Red and Blue casualties are given in the format (xx/xx). Therefore, case A, replication 1 had 112 Red casualties and 24 Blue casualties. The other variable notation has been defined earlier in this paper. This represents the total input required to do the bootstrap experiment and compute the confidence interval estimates using error propagation and Fieller's theorem.

4. RESULTS. Tables 2, 3, 4 and 5 show the results of the bootstrap experiment and the error propagation and Fieller's theorem results for 99%, 95%, 90% and 85% confidence intervals, respectively. The results of the bootstrap method are based on 3000, 1000, ..., 100 replications. The upper limit (UL) and lower limit (LL) are given for the stated level of confidence for all estimates. The average value of the $LER(\hat{LER})$ is given for each bootstrap replication size in addition to the estimates obtained from error propagation and Fieller's theorem. The width of the confidence interval is given as the difference between UL and LL. For example, refer to Table 2. The case A 99% confidence statement of the bootstrap estimate based on 3000 replications is 5.2 for the LER . The upper and lower 99% confidence limits are 6.1 and 4.6, respectively. The width of the confidence interval is 1.5. Fieller's theorem gives upper and lower 99% confidence limits of 6.5 and 4.3 with an interval width of 2.2. Error propagation statistics were 5.2 for the estimate of the LER and 6.2 and 4.1 for the 99% confidence limits.

5. EMERGING TRENDS. The following emerging trends are based on the results shown in tables 2 through 5. These trends should be interpreted with respect to LER s appropriate to the forces and systems modeled using this wargame.

a. The upper and lower confidence limits and the LER estimate ($\hat{LER}$) are relatively insensitive to the replication size (from 3000 to 100) for the four levels of α examined in this study. This was true for all 12 cases (A through L) for the 95, 90 and 85% confidence levels and true for about 3/4 of the cases at the 99% confidence level.

b. The bootstrap confidence interval is consistently shorter than intervals generated by either the error propagation or Fieller's theorem method.

c. Regarding the 99% and 95% confidence intervals, the bootstrap and Fieller's theorem interval estimates tend to yield LER distributions with positive skews. This effect is slightly stronger for the 99% confidence interval than for the 95% confidence interval. This same effect is also true for 90% and 85% interval estimates but not to the same degree as for the 99% and 95% intervals. In fact, the effect is relatively negligible for these two cases.

d. Regarding the 99% confidence interval, the bootstrap lower limit is better approximated by the Fieller's theorem estimates and the bootstrap upper limit is better approximated by the error propagation estimate. Although these findings are relatively consistent across all 12 cases, the degree of agreement is not always good.

e. Regarding the 95% confidence interval, neither the error propagation or Fieller's method has a strong advantage in approximating the bootstrap interval estimates. However, when the error propagation results do a better job in approximating the bootstrap estimates, it generally better approximates the upper confidence limit. The Fieller's theorem method most often approximates the bootstrap lower confidence limit. These 95% findings are consistent with the findings for the 99% confidence interval.

f. Regarding the 90% and 85% confidence intervals, the error propagation and Fieller's theorem estimates are, for the most part, good approximations to the bootstrap results.

TABLE 1. INPUT DATA REQUIRED FOR CASE STUDY
(Based on Force-on-Force Model)

Rep #	CASES:											
	A	B	C	D	E	F	G	H	I	J	K	L
1	112/24	68/34	79/24	99/21	112/15	128/23	99/21	103/26	98/20	56/72	64/22	57/29
2	108/15	72/24	103/21	108/26	105/22	110/21	77/30	120/16	100/17	75/69	70/24	80/24
3	110/17	93/24	95/23	119/18	93/21	85/23	86/28	118/24	105/21	71/55	67/25	88/19
4	103/21	83/30	99/22	104/24	112/23	91/23	88/28	112/24	96/18	73/57	121/21	59/24
5	109/20	56/29	92/23	102/21	92/25	101/23	83/25	116/13	109/20		51/21	74/25
6	112/25	68/31		105/27	98/25	111/17					66/30	17/26
7	100/22	96/23		110/18		105/22					66/26	70/23
8	108/23	70/30		99/25							70/27	53/21
9		74/34		108/17							52/25	71/29
10											68/29	
11											78/25	
12											56/28	
N	8	9	5	9	6	7	5	5	5	4	12	9
R	107.75	75.56	93.60	106.00	102.00	104.43	86.60	113.80	101.60	68.75	69.08	69.89
S _R	4.23	12.85	9.15	6.20	9.01	14.14	8.08	6.72	5.32	8.66	18.13	11.56
B	20.88	28.78	22.60	21.90	21.83	21.71	26.40	20.60	19.20	63.25	25.25	25.11
S _B	3.44	4.21	1.14	3.80	3.71	2.21	3.51	5.73	1.64	8.50	2.96	3.14
R	.11	-.58	-.93	-.47	-.56	-.27	-.84	-.62	.55	-.55	-.31	-.63
μ _{RR}	5.16	2.63	4.14	4.84	4.67	4.81	3.28	5.52	5.29	1.09	2.74	2.78
t _{.99}	3.499	3.355	4.604	3.355	4.032	3.707	4.604	4.604	4.604	5.841	3.106	3.355
t _{.95}	2.365	2.306	2.776	2.306	2.571	2.447	2.776	2.776	2.776	3.182	2.201	2.306
t _{.90}	1.895	1.860	2.132	1.860	2.015	1.943	2.132	2.132	2.132	2.353	1.796	1.860
t _{.85}	1.617	1.592	1.779	1.592	1.699	1.628	1.779	1.779	1.779	1.925	1.549	1.592

TABLE 2. 99% CONFIDENCE INTERVALS FOR LER

CASE		3000	1000	BOOTSTRAP		250	100	FIELLER'S THEOREM	ERROR PROP.
				750	500				
A	UL	6.1	6.2	6.1	6.1	6.0	5.8	6.5	6.2
	LER	5.2	5.2	5.2	5.1	5.2	5.1	(5.2)	5.2
	LL	4.6	4.6	4.6	4.6	4.6	4.7	4.3	4.1
	Width	1.5	1.6	1.5	1.5	1.4	1.1	2.2	2.1
B	UL	3.4	3.3	3.4	3.6	3.4	3.4	3.6	3.5
	LER	2.6	2.6	2.6	2.6	2.7	2.7	(2.6)	2.6
	LL	2.1	2.2	2.1	2.2	2.2	2.2	1.9	1.8
	Width	1.3	1.1	1.3	1.4	1.2	1.2	1.7	1.7
C	UL	4.7	4.7	4.7	4.8	4.8	4.7	5.5	5.4
	LER	4.1	4.1	4.1	4.2	4.2	4.1	(4.1)	4.1
	LL	3.6	3.5	3.6	3.5	3.3	3.5	3.0	2.9
	Width	1.2	1.2	1.1	1.3	1.5	1.2	2.5	2.5
D	UL	5.8	5.8	5.9	5.8	5.9	5.7	6.2	6.0
	LER	4.9	4.9	4.9	4.9	4.9	4.8	(4.8)	4.8
	LL	4.2	4.1	4.2	4.2	4.2	4.3	3.9	3.7
	Width	1.6	1.7	1.7	1.6	1.7	1.4	2.3	2.3
E	UL	6.2	6.1	6.3	5.9	6.1	6.3	7.1	6.4
	LER	4.7	4.7	4.7	4.7	4.7	4.7	(4.7)	4.7
	LL	3.9	3.9	3.8	3.9	3.9	3.9	3.3	2.9
	Width	2.3	2.2	2.5	2.0	2.2	2.4	3.8	3.5
F	UL	5.7	5.7	5.7	5.8	5.8	5.4	6.2	6.1
	LER	4.8	4.8	4.8	4.8	4.8	4.8	(4.8)	4.8
	LL	4.0	4.0	4.1	4.0	4.1	4.2	3.6	3.5
	Width	1.7	1.7	1.6	1.8	1.7	1.2	2.6	2.6
G	UL	4.3	4.4	4.1	4.4	4.0	4.4	5.3	4.7
	LER	3.3	3.3	3.3	3.3	3.3	3.4	(3.4)	3.4
	LL	2.7	2.7	2.7	2.7	2.7	2.7	2.1	1.8
	Width	1.6	1.7	1.4	1.7	1.3	1.7	3.2	2.9
H	UL	8.3	8.0	7.6	8.4	8.0	7.2	13.9	9.1
	LER	5.6	5.6	5.6	5.7	5.6	5.5	(5.5)	5.5
	LL	4.2	4.2	4.3	4.2	4.3	4.1	3.2	1.9
	Width	4.1	3.8	3.3	4.2	3.7	3.1	10.7	7.2
I	UL	5.7	5.8	5.7	5.7	5.7	5.8	6.2	6.1
	LER	5.3	5.3	5.3	5.3	5.3	5.3	(5.3)	5.3
	LL	5.0	4.9	4.9	4.9	5.0	5.0	4.6	4.5
	Width	.7	.9	.8	.8	.7	.8	1.6	1.6
J	UL	1.3	1.3	1.3	1.3	1.3	1.3	2.2	1.8
	LER	1.1	1.1	1.1	1.1	1.1	1.1	(1.1)	1.1
	LL	.8	.8	.8	.9	.9	.9	.5	.4
	Width	.5	.5	.5	.4	.4	.4	1.7	1.4

TABLE 2. 99% CONFIDENCE INTERVALS FOR LER (CONT'D)

<u>CASE</u>		<u>3000</u>	<u>1000</u>	<u>BOOTSTRAP</u>		<u>250</u>	<u>100</u>	<u>FIELLER'S</u> <u>THEOREM</u>	<u>ERROR</u> <u>PROP.</u>
				<u>750</u>	<u>500</u>				
K	UL	3.5	3.5	3.5	3.5	3.6	3.4	3.6	3.5
	LER	2.7	2.7	2.7	2.8	2.7	2.7	(2.7)	2.7
	LL	2.3	2.3	2.3	2.4	2.3	2.3	2.0	2.0
	Width	1.2	1.2	1.2	1.1	1.3	.9	1.6	1.5
L	UL	3.5	3.5	3.5	3.5	3.5	3.5	3.7	3.6
	LER	2.8	2.8	2.8	2.8	2.8	2.8	(2.8)	2.8
	LL	2.3	2.3	2.3	2.3	2.3	2.4	2.1	2.0
	Width	1.2	1.2	1.2	1.2	1.2	1.1	1.6	1.6

TABLE 3. 95% CONFIDENCE INTERVALS FOR LER

CASE		3000	1000	BOOTSTRAP		250	100	FIELLER'S THEOREM	ERROR PROP.
				750	500				
A	UL	5.9	5.8	5.8	5.9	5.8	5.8	5.9	5.9
	LER	5.2	5.2	5.2	5.2	5.2	5.2	(5.2)	5.2
	LL	4.7	4.7	4.7	4.7	4.7	4.7	4.5	4.4
	Width	1.2	1.1	1.1	1.2	1.1	1.1	1.4	1.5
B	UL	3.2	3.1	3.1	3.1	3.2	3.2	3.3	3.2
	LER	2.6	2.6	2.6	2.6	2.6	2.7	(2.6)	2.6
	LL	2.2	2.2	2.2	2.2	2.2	2.3	2.1	2.1
	Width	1.0	.9	.9	.9	1.0	.9	1.2	1.1
C	UL	4.6	4.6	4.6	4.6	4.6	4.6	4.9	4.9
	LER	4.2	4.1	4.1	4.1	4.1	4.2	(4.1)	4.1
	LL	3.7	3.7	3.7	3.7	3.7	3.8	3.4	3.4
	Width	.9	.9	.9	.9	.9	.8	1.5	1.5
D	UL	5.6	5.5	5.6	5.6	5.4	5.4	5.7	5.6
	LER	4.9	4.9	4.9	4.9	4.8	4.9	(4.8)	4.8
	LL	4.3	4.3	4.3	4.3	4.3	4.4	4.2	4.1
	Width	1.3	1.2	1.3	1.3	1.1	1.0	1.5	1.5
E	UL	5.7	5.7	5.6	5.7	5.7	5.8	6.0	5.8
	LER	4.7	4.7	4.7	4.7	4.7	4.6	(4.7)	4.7
	LL	4.0	4.0	4.1	4.0	4.0	4.0	3.7	3.5
	Width	1.7	1.7	1.5	1.7	1.7	1.8	2.3	2.3
F	UL	5.5	5.5	5.5	5.4	5.5	5.4	5.7	5.7
	LER	4.8	4.8	4.8	4.8	4.8	4.8	(4.8)	4.8
	LL	4.2	4.2	4.3	4.2	4.2	4.2	4.0	4.0
	Width	1.3	1.3	1.2	1.2	1.3	1.2	1.7	1.7
G	UL	4.0	3.9	4.0	4.0	4.0	4.0	4.3	4.2
	LER	3.3	3.3	3.3	3.3	3.3	3.3	(3.3)	3.3
	LL	2.8	2.8	2.8	2.8	2.8	2.8	2.5	2.4
	Width	1.2	1.1	1.2	1.2	1.2	1.2	1.9	1.7
H	UL	7.3	7.3	7.4	7.2	7.4	7.4	8.8	7.7
	LER	5.6	5.6	5.6	5.5	5.6	5.5	(5.5)	5.5
	LL	4.4	4.5	4.3	4.3	4.3	4.4	3.9	3.3
	Width	2.9	2.8	3.1	2.9	3.1	3.0	4.9	4.4
I	UL	5.6	5.6	5.6	5.6	5.6	5.7	5.8	5.8
	LER	5.3	5.3	5.3	5.3	5.3	5.3	(5.3)	5.3
	LL	5.0	5.0	5.0	5.0	5.0	5.1	4.9	4.8
	Width	.6	.6	.6	.6	.6	.6	.9	1.0
J	UL	1.3	1.3	1.3	1.3	1.3	1.2	1.6	1.5
	LER	1.1	1.1	1.1	1.1	1.1	1.1	(1.1)	1.1
	LL	.9	.9	.9	.9	.9	.9	.8	.7
	Width	.4	.4	.4	.4	.4	.3	.8	.8

TABLE 3. 95% CONFIDENCE INTERVALS FOR LER (CONT'D)

<u>CASE</u>		<u>3000</u>	<u>1000</u>	<u>BOOTSTRAP</u>		<u>250</u>	<u>100</u>	<u>FIELLER'S THEOREM</u>	<u>ERROR PROP.</u>
				<u>750</u>	<u>500</u>				
K	UL	3.3	3.3	3.3	3.3	3.3	3.2	3.3	3.3
	LER	2.7	2.7	2.7	2.7	2.7	2.7	(2.7)	2.7
	LL	2.4	2.4	2.4	2.4	2.4	2.4	2.2	2.2
	Width	.9	.9	.9	.9	.9	.8	1.1	1.1
L	UL	3.3	3.3	3.3	3.3	3.3	3.2	3.4	3.3
	LER	2.8	2.8	2.8	2.8	2.8	2.8	(2.8)	2.8
	LL	2.4	2.3	2.4	2.4	2.3	2.5	2.3	2.2
	Width	.9	1.0	.9	.9	1.0	.7	1.1	1.1

TABLE 4. 90% CONFIDENCE INTERVALS FOR LER

CASE		BOOTSTRAP						FIELLER'S THEOREM	ERROR PROP.
		3000	1000	750	500	250	100		
A	UL	5.7	5.7	5.7	5.7	5.7	5.7	5.8	5.7
	LER	5.2	5.2	5.2	5.2	5.2	5.2	(5.2)	5.2
	LL	4.8	4.8	4.8	4.8	4.7	4.8	4.6	4.6
	Width	.9	.9	.9	.9	1.0	.9	1.2	1.1
B	UL	3.1	3.0	3.1	3.1	3.0	3.1	3.1	3.1
	LER	2.6	2.6	2.6	2.6	2.6	2.7	(2.6)	2.6
	LL	2.3	2.3	2.3	2.3	2.3	2.3	2.2	2.2
	Width	.8	.7	.8	.8	.7	.8	.9	.9
C	UL	4.5	4.6	4.5	4.5	4.6	4.6	4.7	4.7
	LER	4.1	4.1	4.1	4.1	4.2	4.1	(4.1)	4.1
	LL	3.7	3.7	3.7	3.7	3.8	3.8	3.6	3.6
	Width	.8	.9	.8	.8	.8	.8	1.1	1.1
D	UL	5.4	5.4	5.5	5.5	5.4	5.3	5.5	5.5
	LER	4.9	4.9	4.9	4.8	4.8	4.9	(4.8)	4.8
	LL	4.4	4.4	4.4	4.4	4.4	4.4	4.3	4.2
	Width	1.0	1.0	1.1	1.1	1.0	.9	1.2	1.3
E	UL	5.5	5.4	5.4	5.3	5.5	5.5	5.7	5.6
	LER	4.7	4.7	4.7	4.7	4.7	4.7	(4.7)	4.7
	LL	4.1	4.1	4.1	4.1	4.1	4.2	3.9	3.8
	Width	1.4	1.3	1.3	1.2	1.4	1.3	1.8	1.8
F	UL	5.4	5.4	5.4	5.4	5.3	5.3	5.5	5.5
	LER	4.8	4.8	4.8	4.8	4.8	4.8	(4.8)	4.8
	LL	4.3	4.3	4.3	4.3	4.3	4.3	4.2	4.1
	Width	1.1	1.1	1.1	1.1	1.0	1.0	1.3	1.4
G	UL	3.8	3.8	3.8	4.0	3.8	3.9	4.0	4.0
	LER	3.3	3.3	3.3	3.3	3.3	3.3	(3.3)	3.3
	LL	2.9	2.9	2.9	2.9	2.9	2.9	2.7	2.6
	Width	.9	.9	.9	1.1	.9	1.0	1.3	1.4
H	UL	7.1	7.1	7.1	7.1	6.9	7.1	7.8	7.2
	LER	5.6	5.6	5.6	5.6	5.5	5.6	(5.6)	5.6
	LL	4.5	4.5	4.5	4.6	4.5	4.5	4.2	3.8
	Width	2.6	2.6	2.6	2.6	2.4	2.6	3.6	3.4
I	UL	5.6	5.6	5.6	5.6	5.6	5.6	5.7	5.7
	LER	5.3	5.3	5.3	5.3	5.3	5.3	(5.3)	5.3
	LL	5.1	5.1	5.1	5.1	5.1	5.1	5.0	4.9
	Width	.5	.5	.5	.5	.5	.5	.7	.8
J	UL	1.3	1.2	1.3	1.2	1.3	1.3	1.4	1.4
	LER	1.1	1.1	1.1	1.1	1.1	1.1	(1.1)	1.1
	LL	.9	.9	.9	.9	.9	.9	.8	.8
	Width	.4	.3	.4	.3	.4	.4	.6	.6

TABLE 4. 90% CONFIDENCE INTERVALS FOR LER (CONT'D)

<u>CASE</u>		<u>3000</u>	<u>1000</u>	<u>BOOTSTRAP</u>		<u>250</u>	<u>100</u>	<u>FIELLER'S THEOREM</u>	<u>ERROR PROP.</u>
				<u>750</u>	<u>500</u>				
K	UL	3.2	3.2	3.2	3.2	3.2	3.2	3.2	3.2
	LER	2.8	2.8	2.7	2.7	2.7	2.7	(2.7)	2.7
	LL	2.4	2.4	2.4	2.4	2.4	2.4	2.3	2.3
	Width	.8	.8	.8	.8	.8	.8	.9	.9
L	UL	3.2	3.2	3.2	3.2	3.1	3.2	3.3	3.2
	LER	2.8	2.8	2.8	2.8	2.8	2.8	(2.8)	2.8
	LL	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.3
	Width	.8	.8	.8	.8	.7	.8	.9	.9

TABLE 5. 85% CONFIDENCE INTERVALS FOR LER

CASE		3000	1000	BOOTSTRAP		250	100	FIELLER'S THEOREM	ERROR PROP.
				750	500				
A	UL	5.6	5.6	5.6	5.6	5.7	5.6	5.7	5.6
	LER	5.2	5.2	5.2	5.2	5.2	5.2	(5.2)	5.2
	LL	4.8	4.8	4.8	4.8	4.8	4.9	4.7	4.7
	Width	.8	.8	.8	.8	.9	.7	1.0	.9
B	UL	3.0	3.0	3.0	3.0	3.0	3.0	3.0	3.0
	LER	2.6	2.6	2.6	2.7	2.7	2.6	(2.6)	2.6
	LL	2.3	2.3	2.3	2.3	2.3	2.3	2.3	2.2
	Width	.7	.7	.7	.7	.7	.7	.7	.8
C	UL	4.5	4.5	4.5	4.5	4.5	4.5	4.6	4.6
	LER	4.2	4.2	4.1	4.1	4.2	4.1	(4.1)	4.1
	LL	3.8	3.8	3.8	3.8	3.8	3.8	3.7	3.7
	Width	.7	.7	.7	.7	.7	.7	.9	.9
D	UL	5.3	5.3	5.3	5.4	5.3	5.3	5.4	5.4
	LER	4.9	4.9	4.9	4.9	4.9	4.9	(4.8)	4.8
	LL	4.4	4.4	4.4	4.4	4.5	4.5	4.4	4.3
	Width	.9	.9	.9	1.0	.8	.8	1.0	1.1
E	UL	5.3	5.3	5.3	5.4	5.3	5.4	5.5	5.4
	LER	4.7	4.7	4.7	4.7	4.7	4.7	(4.7)	4.7
	LL	4.2	4.1	4.1	4.2	4.1	4.2	4.0	3.9
	Width	1.1	1.2	1.2	1.2	1.2	1.2	1.5	1.5
F	UL	5.3	5.3	5.3	5.3	5.2	5.3	5.4	5.4
	LER	4.8	4.8	4.8	4.8	4.8	4.8	(4.8)	4.8
	LL	4.4	4.3	4.4	4.4	4.4	4.3	4.3	4.2
	Width	.9	1.0	.9	.9	.8	1.0	1.1	1.2
G	UL	3.7	3.7	3.7	3.7	3.7	3.7	3.9	3.8
	LER	3.3	3.3	3.3	3.3	3.3	3.3	(3.3)	3.3
	LL	2.9	2.9	2.9	2.9	2.9	2.9	2.8	2.7
	Width	.8	.8	.8	.8	.8	.8	1.1	1.1
H	UL	6.7	6.9	6.9	6.6	6.9	6.6	7.3	6.9
	LER	5.6	5.6	5.6	5.6	5.7	5.5	(5.5)	5.5
	LL	4.7	4.7	4.7	4.6	4.7	4.5	4.4	4.1
	Width	2.0	2.2	2.2	2.0	2.2	2.1	2.9	2.8
I	UL	5.5	5.5	5.6	5.5	5.5	5.5	5.6	5.6
	LER	5.3	5.3	5.3	5.3	5.3	5.3	(5.3)	5.3
	LL	5.1	5.1	5.1	5.1	5.1	5.1	5.0	5.0
	Width	.4	.4	.5	.4	.4	.4	.6	.6
J	UL	1.2	1.2	1.2	1.2	1.2	1.2	1.4	1.3
	LER	1.1	1.1	1.1	1.1	1.1	1.1	(1.1)	1.1
	LL	.9	.9	.9	1.0	1.0	.9	.9	.9
	Width	.3	.3	.3	.2	.2	.3	.5	.4

TABLE 5. 85% CONFIDENCE INTERVALS FOR LER (CONT'D)

<u>CASE</u>		<u>3000</u>	<u>1000</u>	<u>BOOTSTRAP</u>		<u>250</u>	<u>100</u>	<u>FIELLER'S THEOREM</u>	<u>ERROR PROP.</u>
				<u>750</u>	<u>500</u>				
K	UL	3.1	3.1	3.1	3.1	3.1	3.1	3.1	3.1
	LER	2.7	2.7	2.7	2.7	2.7	2.7	(2.7)	2.7
	LL	2.4	2.4	2.4	2.4	2.4	2.4	2.4	2.3
	Width	.7	.7	.7	.7	.7	.7	.7	.8
L	UL	3.1	3.1	3.1	3.1	3.1	3.1	3.2	3.2
	LER	2.8	2.8	2.8	2.8	2.8	2.8	(2.8)	2.8
	LL	2.5	2.5	2.5	2.5	2.5	2.5	2.4	2.4
	Width	.6	.6	.6	.6	.6	.6	.8	.8

Postscript submitted by
Joseph M. Tessmer
1947th HQ Support Group
SAGF
Directorate for Theater Force Analysis
Fighter Division
USAF
Washington, DC 20330

Subject: Unusual Data Sets

The following results were obtained from a force-on-force evaluation.

Replication	Red Casualties	Blue Casualties
1	.9687	0
2	.5069	0
3	.5086	0
4	.1362	.3274
5	0	0
6	.1405	0
7	0	0

Note that the number of Blue casualties is zero six times out of seven and both Red and Blue casualties are zero two times out of seven. The application of equation (5), Fieller's theorem, yields 90% confidence limits of -7.41 and .62. The estimate of the LER (equation 2) is 6.93. In this case, the upper and lower confidence limits do not include the point estimate of the LER. The results of the bootstrap are also seemingly anomalous. The mean LER value is about 1200 across different replication sizes ranging from 250 through 3000 and the upper and lower 90% confidence limits average about 7000 and 2.5, respectively. This observation does not negate the use of the bootstrap and Fieller's method, but does indicate that some unusual data sets (i.e., containing a preponderance of zeros and numbers less than one) should not be analyzed in this fashion. More theoretical work needs to be done concerning the make-up of the data before subjecting them to analysis. Of course, this is true for any statistical procedure.

REFERENCES

1. Diaconis, Persie; Efron, Bradley; Computer Intensive Methods in Statistics; Scientific American, May 1983.
2. Dutoit, E.; Estimating the Variance of the Loss Exchange Ratio; Proceedings of the Twenty-Eighth Conference on the Design of Experiments in Army Research Development and Testing, ARO Report 83-2, October 1982.

The following references contain useful information concerning error propagation and Fieller's theorem.

- a. Beers, Y; Introduction to the Theory of Error, Addison-Wesley, 1957.
- b. Finney, D.J.; Statistical Method in Biological Assay, Charles Griffen and Co limited.
- c. Fishman, G; Principals of Discrete Event Simulation, Wiley, 1978.
- d. Goldstein, A; Biostatistics, An Introductory Text, Mac Millan Company, 1968.
- e. Rao, C.R.; Linear Statistical Inference and Its Applications, Wiley, 1965.

ACCEPTANCE OF A MEAL AND ITS COMPONENTS - AN EXERCISE IN MISSING DATA

Edward W. Ross
Staff Mathematician
US Army Natick R&D Laboratories

ABSTRACT

This paper is a study of the relation between the consumer acceptance of a meal and of the items that make up the meal. The primary purpose is to find a way of predicting overall meal-acceptance scores from the scores for the individual items in the Army field-ration system called the Meal, Ready-to-Eat. Attempts to do a linear regression encounter difficulties because of the large and non-random fraction of missing data. This problem is treated by a procedure that leads eventually to a single formula for the predicted overall meal scores. These predicted meal scores are then analyzed by the same methods used for the item scores. Stability results for meals are found using data from a storage study of the items after 24 months.

Introduction

This paper describes the study of the relationship between the food items in a meal and the meal considered as a whole, in terms of the scores which the item and meal receive in a consumer-acceptance test. The purpose is to derive and apply a formula that will permit estimation of meal scores from scores of the items in the meal.

This effort has its origins in a storage study of a military ration system called the Meal, Ready-to-Eat (MRE) which is now in progress at the U. S. Army Natick R&D Laboratories. In this study consumers are asked to evaluate the items in a meal but do not give an evaluation of the overall meal. However, when the meal is used in the field, it will be judged as a whole. Consequently, it is desirable to have a way of estimating the acceptance score for a meal from the scores of the items in the meal. Such an algorithm allows one to study how meal-acceptance is affected by storage time and temperature.

Previous work on this question is described in a report by Rogozenski and Moskowitz (1974), which also mentions earlier efforts in this direction. Their principal finding was that meal score was governed primarily by the entree score; the other meal components (starch, vegetable, salad and dessert) had less than one-third as much influence as the entree. This accords well with intuition. Their procedure was, as ours will be, mainly a statistical regression analysis of a large set of data on items and meals. The military

meals used in their analysis were typical of those served in a garrison setting, i.e. a mess hall, which are much different from combat rations like the MRE. The present study differs from theirs also in the important role played by the treatment of missing data in the analysis.

Materials and Methods

In this section we describe the storage study, then discuss the resulting data and finally present the procedure for predicting meal scores and analyzing them to find their storage-stability.

Sketch of the Storage Study

The storage study of the MRE ration is described by Ross et al (1983). We give here a brief summary of this investigation.

The MRE consists of 12 meals or menus, each containig roughly six items, not all of which are evaluated in this test. There are 39 different items in all, a number of which occur in more than one meal. Each menu contains an entree plus items of other types. These types are as follows:

- type 1 - entrees
- 2 - pastries
- 3 - vegetables
- 4 - fruits
- 5 - spreads
- 6 - beverages

7 - candies

8 - miscellaneous (catsup, crackers, etc.)

In the storage study these meals were obtained through the usual Defense Department procurement system. Some of the meals were tested when received, and the rest stored at temperatures of 4, 21, 30 and 38 degrees C. The meals were withdrawn from storage and served to test subjects according to the schedule shown in Table 1. In these tests each of 36 consumers evaluates at one sitting all the items in a meal, assigning to each a score on the 9-point hedonic scale

9 means "like extremely"

.

.

5 means "neither like nor dislike"

.

.

1 means "dislike extremely"

After a withdrawal the scores for each item at all the preceding and current withdrawals are analyzed by a variety of statistical tests to estimate their shelf-lives and various other characteristics of their storage stability.

Ordinarily each item is analyzed apart from all others. Indeed, if an item is present in more than one meal, it is studied as a separate item in each meal where it occurs. However, at the withdrawal following 24 months of storage test subjects were asked to furnish evaluations of the meals as a whole in addition to, and on the same scale as, the items in the meal. We use these data to develop a model for predicting the overall meal scores. Thereafter the model is applied to data at other withdrawal-times to predict meal scores, and

the meal scores for all times and temperatures for which we have data are then analyzed by the same routines used for the individual items.

Data Structure

The data on which the model is based are item-scores and meal-scores, obtained after 24 months of storage at three temperatures, 21, 30 and 38 deg. C. (Table 1 shows that meals stored at 4 deg C. were not tested at the 24-month withdrawal). For each menu at each of the three storage temperatures the data were placed in an array of 36 rows and 9 columns, a row containing the scores given by a test-subject, and the column designating the food type, 1 through 8. Column 9 contains the meal scores. Each meal includes items of certain types and not of others. The symbol 0 (zero) is used as a missing-data indicator and appears in columns for food-types absent from a meal.

The description of the data is furthered by Tables 2 and 3. Table 2 lists each of the 52 items by name and gives its item-index, the index of the menu in which it occurs and the food-type. Table 3 is an array showing for each menu the food-types present and the indices of the items. E.g. we see that Menu 10 includes items of Type 1 (Item 10, meatballs), Type 2 (Item 22, chocolate nut cake), Type 3 (Item 28, potato patty) and Type 8 (Item 52, crackers and jelly), but no items of Types 4, 5, 6 nor 7. Table 2 makes clear that several foods (brownies, cookies, etc) occur as different items in different menus. Table 3 shows that entrees occur in every menu, but the other types do not.

We shall see in the next sub-section that the estimation and prediction is based (at least initially) on the following model of linear regression:

$$X_{ij} = \sum_{j=1}^8 X_{ij} B_j + e_i \quad i=1,2,\dots,36 \quad (1)$$

Here X_{ij} is the score given for the item of type j by the i -th respondent, and $j = 9$ denotes the meal score. The B 's are the regression-coefficients to be estimated, the e 's are assumed to be independent, Gaussian random variables with

$$\text{mean}(e) = 0 \quad (2)$$

$$\text{stddev}(e) = S \quad (3)$$

and S is also to be estimated. We are, therefore, attempting to fit the vector of meal-scores by a linear combination of vectors of type-scores. Any line of data which lacks one or more type-scores is incomplete and will be classified as missing data by most of the common computer algorithms for treating data. We see from Table 3 that every menu, and hence every line of data, has some missing types, and so our entire data-set will be classified as missing! This suggests that our way of handling missing-data will have an important effect on the results of the regression.

If we examine this data-set, we are struck by the inherent nature of the missing data as well as its prevalence. Ordinarily, when data are missing in an experiment, it is accidental and occurs in relatively few cases. In the present context, it is by no means

accidental; on the contrary, it results from the desire to introduce variety into the meals and is completely intentional. Also, roughly 45% of the data-cells are missing, see Table 3, hardly a small fraction of the total data.

Either of these circumstances is sufficient to rule out successful application of the usual procedure for handling missing data, the EM-algorithm, see Dempster et al (1977) and Laird and Louis (1982). Some other approach is needed in such problems of structural missing data.

Moreover, the situation is in fact somewhat worse than so far depicted for several reasons. First, the data at 24 months were censored in the following way. The test-monitors decided that the following items were unfit for consumption by the test-subjects:

Strawberries, Items 31 and 32 at 38 degree storage

Brownies, Items 13 and 14, at 21, 30, and 38 degrees

Data for these cases also appear as missing in the data array.

Second, the test monitors forgot to ask for overall meal scores for Menus 1 and 2 at 24 months from 21 and 30 degree storage although scores were obtained for the individual items in those meals. Finally, there was one instance of more-or-less ordinary missing data, i.e. in Menu 7 for 30 storage only 35 lines of data were taken.

Thus we have missing data for a variety of causes. The situation is depicted in the missing-data map, Table 4, which shows for each menu and storage temperature whether the various food-types have

missing data, and, if so, the reason. In this chart blank entries imply that we have data, the symbol S denotes structural or inherent missing data, C indicates censored data, F the forgotten cases and numerical entries specify the number of missing scores due to unexplained or random mishaps. The number of missing scores is 36 (i.e. all scores) in the cases marked by letter symbols. To reiterate, the extent and complexity of the missing data have a major effect on the estimation and prediction procedures that we use.

Estimation and Prediction

The fact that the data consist of discrete scores suggests that we could do either a regression analysis or one based on some form of contingency tables. We choose the former because its methods are more completely available in computer software but remain cautious about the applicability of its underlying assumptions.

If missing data were not a problem, we would do a stepwise regression, bringing in one food-type at a time and ceasing when additional food-types caused no improvement in the fit. If we attempted this in the present situation, we would find that bringing in a new type would perhaps improve the fit but would usually also reduce the number of data cases on which the estimates are based. Eventually, bringing in a new food type would have no effect because all the data would be classified as missing. For example, any regression involving both Type 3 (vegetables) and Type 4 fruit) would have missing data in every line, see Table 3.

In deciding how to proceed, we list the characteristics that we would like the final method to have, keeping in mind that we intend to use the results for predicting meal scores where none have been measured.

1. The fit should be good.
2. The fit should be based on as large a data-set as possible.
3. The fit should be free of peculiarities caused by correlated columns (near rank-deficiency) or excessively influential data points.
4. The resulting predictor should use only item scores that are available.
5. The predictor should be as simple as possible.

Clearly these desires conflict, and we must seek a compromise among them. Many different procedures are possible, some which produce an excellent fit for a small data-set, others a poorer fit applicable to a large data-set etc.

We describe now the procedure that was finally used. It is based on (a) pooling food-types and (b) estimating missing data from entree scores. To be precise we introduce a vector, Y , of transformed scores as follows:

$$\begin{aligned}
 Y_{ij} &= X_{ij} & j=1,2 \\
 Y_{13} &= X_{13} + X_{14}, & Y_{i4} &= X_{i5} + X_{i8} \\
 Y_{15} &= X_{16}, & Y_{i6} &= X_{i7}
 \end{aligned} \tag{4}$$

In these formulas we explicitly use the missing-data symbol, i.e. the X value is taken as 0 if the data is missing. This transformation amounts to pooling types 3 and 4 (vegetables and fruits) into a new type, 3, and similarly pooling Types 5 and 8 (spreads and miscellaneous) into another new type, 4, and re-indexing Types 5 and 6 to avoid confusion. The reasons for these choices are visible in Table 3. Types 3 and 4 are almost perfectly complementary in the sense that one has data where the other lacks it, and there are no cases where both have data. Likewise for Types 5 and 8. Types 6 and 7 lack these desirable properties. Moreover the pooling of scores for fruits and vegetables makes some sense from a food-technological viewpoint since both are plant products. The pooling of spreads and miscellaneous types lacks as clear a justification, but we argue that the original classification of these types is somewhat arbitrary, and the present pooling is no more so.

The pooling creates a new set of types which show much less missing data than the original classification. The Y-variables numbered 2, 3 and 4 are each missing from only one menu, respectively 6, 9 and 12. We estimate the missing scores for each of these cases by regressing the variable on the ubiquitous variable for entrees, number 1, over the 11 meals where data are present. Then that relation is used to predict the scores for the lone missing case. We find

$$Y_2 = 5.65 + 0.172Y_1 \quad (r = 0.21) \quad (5)$$

$$Y_3 = 4.77 + 0.282Y_1 \quad (r = 0.33) \quad (6)$$

$$Y_4 = 4.17 + 0.326Y_1 \quad (r = 0.34) \quad (7)$$

In all cases the t-values for the coefficients exceed 6.

With these procedures we obtain a set of data for Y-variables number 1,2,3,4 and 9 that is complete except for the one randomly missing data-point of meal 7 and the forgotten sets of meals 1 and 2. That is, we have filled in the censored and structurally missing data.

This rather large data-set (1151 values) is used as the basis for a linear regression with model

$$X_{19} = B_0 + \sum_{j=1}^4 (Y_{1j} B_j) + e_i \quad (8)$$

The resulting estimates and their t-values are

$$\begin{array}{ll} B_0 = -0.684 & T_0 = -4.25 \\ B_1 = 0.398 & T_1 = 26.8 \\ B_2 = 0.217 & T_2 = 10.4 \\ B_3 = 0.257 & T_3 = 14.9 \\ B_4 = 0.234 & T_4 = 13.7 \end{array} \quad (9)$$

The regression has $r = .838$, $S = 0.920$, and the residuals are

distributed in reasonably Gaussian fashion. The Y-variables 5 and 6 do not make a major improvement to the fit.

The purpose in deriving these estimates is to generate predicted menu scores which can then be subjected to the same storage-stability calculations as the individual items. These algorithms require integer values as input, so the predicted scores obtained from the regression must be rounded to whole numbers. When this is done, and the predicted integers compared with the meal-scores of the data, we find that their differences are distributed as follows:

Difference	Number of Values
-4	3
-3	7
-2	41
-1	208
0	621
1	226
2	40
3	6

The predictor obtained from the regression and missing-data procedure can be written

$$\begin{aligned}
X_9 = & -0.684 + .398Y_1 + 0.217\{U_2 Y_2 + (1-U_2)(5.65 + 0.172Y_1)\} \\
& + 0.257\{U_3 Y_3 + (1-U_3)(4.77 + 0.282Y_1)\} \\
& + 0.234\{U_4 Y_4 + (1-U_4)(4.17 + 0.326Y_1)\} \quad (10)
\end{aligned}$$

where $U_j = 0$ if $Y_j = 0$ (i.e. data for Y_j is missing)

$U_j = 1$ if $Y_j > 0$ (i.e. data for Y_j is present)

and X_9 is rounded to the nearest integer after the calculation.

This predictor is used on the data through 24 months from the MRE to create for each item and temperature a set of predicted scores in exactly the same form as for an item. These data are then run through the calculations that produce estimates of storage stability at the various temperatures.

In predicting meal scores for the cases where censored data occur (items at various temperatures in menus 2, 3, 8 and 12) there is an element of uncertainty in the above procedure. The predictor (10) was derived by omitting these items, i.e. regarding them as missing data. To be perfectly consistent, prediction should be done in the same way. However, we do in fact know that the test-monitors thought that the items were too bad to be served, which implies that the scores would have been very low had the items been tested. Accordingly, the scores for these items were set to 1, the lowest possible score, prior to using (10) to estimate the scores of the menus in which they occur.

Results and Discussion

Table 5 shows estimates of shelf-lives for each menu when stored at the four test-temperatures, based on the meal data through 24 months generated by the procedure described in the preceding section. The shelf-lives in Table 5 are the shortest of three found by doing, respectively, linear and non-linear regressions and a multinomial logit fit for the time-dependence of the data. The procedures are described by Ross et al (1983). Since these are based on data only through 24 months, large estimates of shelf-life are likely to be erratic; consequently, we do not enter shelf-lives predicted to be longer than 48 months in this table. We see that all the meals had shelf lives exceeding 48 months at 4 degrees, as did most of the meals at 21 and 30. At 38 degrees half the meals had lives less than 36 months, the shortest being 19 months for Meal 2. The most stable menus appear to be 1,5,6,7 and 10 though 4 and 9 are also quite good.

Another way of looking at storage-stability is by means of the average scores after some fixed storage time, say 24 months. These are listed in Table 6. In a coarse way we expect that menus with short shelf-lives will have low scores, and this effect can be seen by comparing Tables 5 and 6.

In general the least stable menus appear to be 3 and 8. Much of their instability seems to be caused by imputing scores of 1 to the censored data for brownies, Items 13 and 14, which are part of these meals. For, when we repeat the prediction with brownies excluded, i.e.

treated as missing data, we find that these meals show little or no change over time at any of the temperatures. A similar effect is found for strawberries, Items 31 and 32, in Menus 2 and 12 at 38 degrees although the effect is weaker in Menu 2 than for the others.

It appears then, that there can be a sizeable difference in the time-behavior of menu-scores resulting from the two ways of treating the censored data for an individual item in a menu. In both procedures the censored data for the item is regarded as missing when the regression (8) is done, i.e. artificial scores for the missing item-type are calculated, using Equations (5) or (6), as appropriate. The difference arises in the prediction stage where Equation (10) is used to generate a menu-score from the item-scores. If the scores for an item are thought to be missing, then $U = 0$, and the artificial score for the item is used. If we think that the item scores are 1, then we are taking them as known, $U = 1$, and the 1's rather than the artificial scores are used.

It is not surprising that these procedures lead to different outcomes, and it is also possible to suggest further methods that will give still other results. For example, perhaps assigning all 1's is too extreme. Some different distribution of low scores might be more plausible and lead to less abruptly different results. What is lacking is a rationale for choosing the most plausible distribution. It is clear, however, that the censored data are not randomly missing. We know something though not everything about what these scores would have been had the items been served.

It is also possible to imagine a large number of alternative procedures for handling structural missing data. For example, we could simply do a separate regression for each menu. In that case there is no missing data, but we have 12 different regressions, each based on only 108 scores. Across all menus this leads to a slightly better fit than the procedure in Materials and Methods. Out of 1151 scores there were 85, as opposed to 97, with absolute differences exceeding 1, an improvement that is not significant at the 90% confidence level according to the chi-square test. The procedure in Materials and Methods is at the opposite extreme from the one just given, for we have only a single regression but have had to deal extensively with missing data.

It is also worth inquiring whether the regression results, (9), have been adversely affected by correlations among the variables. There are significantly non-zero correlations among the Y-variables, for we used them in obtaining the results (5) to (7). However, the relations depicted there seem too weak to cause much difficulty with the solution of the normal regression equations. This is confirmed by an eigenvalue analysis of the correlation matrix, which shows no indication of ill-conditioning.

The results shown in Table 5 are not strictly comparable with those found earlier for the MRE by Ross et al (1983) and Ross (1983a). However, in general we might expect the meal results to resemble those for all items pooled in the earlier papers, and they do. For example, in (1983) all items pooled were found to have a shelf-life of 42

months at 38 degrees. Again, in (1983a) a time-temperature model led to an estimated shelf-life for all items of 44 months. The median of the shelf lives for the 12 meals at 38 in Table 5 is 36.5 months. Both the previous efforts suffer from large enough standard errors so the present results are not inconsistent with them.

To conclude, the procedure in Materials and Methods appears to be a plausible one for dealing with problems suffering from large amounts of structural or censored missing data. It is reasonably simple to use and gives results that seem to be sensible. Tentatively, we adopt it for current use and expect to test it against measured data at some future withdrawal.

REFERENCES

- Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977), "Maximum likelihood from incomplete data via the E.M. algorithm," *Journal of the Royal Statistical Society, Ser. B*, 1-38.
- Laird, N.M. and Louis, T.A. (1982), "Approximate posterior distributions from incomplete data problems," *Journal of the Royal Statistical Society, Ser. B*, 190-200.
- Rogozenski, J.E. and Moskowitz, H.R. (1974), "A system for the preference evaluation of cyclic menus," *US Army Natick Laboratories Technical Report 75-46-OR/SA*.
- Ross, E.W., Klicka, M.V., Kalick, J. and Branagan, M.E., "Acceptance of a Military Ration after Twenty-four Month Storage", submitted to *Journal of Food Science, IFT*, (1983).

TABLE 1

TEST SCHEDULE. AN * DENOTES A TEST NOT MADE.

STORAGE TEMP STORAGE DURATION (MONTHS)

4 DEG C	0	*	12	*	*	30	36	48	60	108	*	*
21 DEG C	0	*	12	18	24	30	36	48	60	*	*	126
30 DEG C	0	6	12	18	24	30	36	48	60	*	*	*
38 DEG C	0	6	12	18	24	*	*	*	*	*	114	*

TABLE 2

LIST OF FOOD-ITEMS, THEIR INDICES, MENUS AND TYPES

ITEM: DESCRIPTION	INDEX	MENU	TYPE	ITEM: DESCRIPTION	INDEX	MENU	TYPE
PORK SAUSAGE PATTY	1	1	1	STRAWBERRIES	31	2	4
HAM CHICKEN LOAF	2	2	1	STRAWBERRIES	32	12	4
BEEF PATTY	3	3	1	APPLE SAUCE	33	1	4
BEEF WITH BBQ-SAUCE	4	4	1	FRUIT MIX	34	5	4
BEEF STEW	5	5	1	CHEESE SPREAD	35	3	5
FRANKFURTERS	6	6	1	PEANUT BUTTER	36	4	5
TURKEY WITH GRAVY	7	7	1	JELLY	37	7	5
BEEF WITH GRAVY	8	8	1	COCOA	38	1	6
CHICKEN A LA KING	9	9	1	COCOA	39	7	6
MEATBALLS	10	10	1	COCOA	40	9	6
HAMSLICES	11	11	1	COFFEE	41	12	6
BEEF & SPICE SAUCE	12	12	1	CHOCOLATE TOFFEE	42	4	7
BROWNIES	13	3	2	CHOCOLATE FUDGE	43	6	7
BROWNIES	14	8	2	VANILLA CREME	44	12	7
COOKIES	15	1	2	CATSUP	45	1	8
COOKIES	16	4	2	CATSUP	46	2	8
COOKIES	17	12	2	CRACKERS	47	2	8
PINEAPPLE NUT CAKE	18	2	2	CRACKERS	48	11	8
CHERRY NUT CAKE	19	5	2	CRACKERS & PNT BTR	49	3	8
MAPLE NUT CAKE	20	7	2	CRACKERS & CHEESE	50	8	8
FRUIT CAKE	21	9	2	CRACKERS & CHEESE	51	9	8
CHOCOLATE NUT CAKE	22	10	2	CRACKERS & JELLY	52	10	8
ORANGE NUT CAKE	23	11	2				
BEANS & TOMATO SAUCE	24	3	3				
BEANS & TOMATO SAUCE	25	6	3				
BEANS & TOMATO SAUCE	26	8	3				
POTATO PATTY	27	7	3				
POTATO PATTY	28	10	3				
PEACHES	29	4	4				
PEACHES	30	11	4				

TABLE 3

ITEMS AND TYPES PRESENT IN VARIOUS MEALS. IF AN
ITEM OF A CERTAIN TYPE IS PRESENT IN A MENU, ITS
ITEM-INDEX OCCUPIES THE CORRESPONDING CELL IN THE TABLE

MENU NO.// TYPE NO.:	1	2	3	4	5	6	7	8
-----	--	--	--	--	--	--	--	--
1	1	15		33		38		45
2	2	18		31				47
3	3	13	24		35			
4	4	16		29	36		42	
5	5	19		34				49
6	6		25				43	46
7	7	20	27		37	39		
8	8	14	26					50
9	9	21				40		51
10	10	22	28					52
11	11	23		30				48
12	12	17		32		41	44	

TABLE 4

MISSING DATA MAP. A SYMBOL IN A CELL MEANS
THAT DATA IS MISSING FROM THAT CELL. SEE
TEXT FOR EXPLANATION OF CODE

MENU	TEMP//TYPE	1	2	3	4	5	6	7	8	9
1	21			S		S		S		F
1	30			S		S		S		F
1	38			S		S		S		F
2	21			S		S	S	S		F
2	30			S		S	S	S		F
2	38			S		S	S	S		F
3	21			S	C	S		S		
3	30			C		S	S	S	S	
3	38			C		S	S	S	S	
4	21			S		S		S	S	
4	30			S		S		S	S	
4	38			S		S		S	S	
5	21			S		S	S	S	S	
5	30			S		S	S	S	S	
5	38			S		S	S	S	S	
6	21			S		S	S	S	S	
6	30			S		S	S	S	S	
6	38			S		S	S	S	S	
7	21			S		S	S	S	S	
7	30			S		S	S	S	S	
7	38			S		S	S	S	S	
8	21			C		S	S	S	S	
8	30			C		S	S	S	S	
8	38			C		S	S	S	S	
9	21			S		S	S	S	S	
9	30			S		S	S	S	S	
9	38			S		S	S	S	S	
10	21			S		S	S	S	S	
10	30			S		S	S	S	S	
10	38			S		S	S	S	S	
11	21			S		S	S	S	S	
11	30			S		S	S	S	S	
11	38			S		S	S	S	S	
12	21			S		S		S	S	
12	30			S		S		S	S	
12	38			S	C	S		S	S	

TABLE 5

PREDICTED SHELF-LIVES IN MONTHS OF THE
TWELVE MEALS AT FOUR TEMPERATURES. A -
MEANS THAT THE SHELF-LIFE EXCEEDS 48 MONTHS.

MEAL	4	21	30	38
*****	****	****	****	****
1	-	-	-	-
2	-	-	-	19
3	-	22	26	25
4	-	-	-	39
5	-	-	-	-
6	-	-	-	-
7	-	-	-	-
8	-	-	21	27
9	-	-	-	34
10	-	-	-	-
11	-	-	35	34
12	-	-	-	22

TABLE 6

AVERAGE SCORES OF THE TWELVE MEALS AFTER
 TWENTY-FOUR MONTH STORAGE AT FOUR TEMPERATURES.

MEAL	4	21	30	38
*****	*****	*****	*****	*****
1	6.17	6.76	6.86	6.08
2	6.58	6.17	5.86	4.28
3	5.92	4.42	5.08	4.89
4	6.64	6.80	6.17	5.83
5	6.69	6.83	6.25	6.19
6	5.64	5.78	5.42	5.63
7	6.51	6.49	6.25	6.36
8	5.74	5.03	4.44	5.06
9	6.82	6.82	6.25	5.64
10	6.32	6.33	6.21	5.69
11	6.31	6.94	6.03	5.53
12	6.08	6.25	6.27	4.53

NUMERICAL VALIDATION OF TUKEY'S CRITERION FOR CLINICAL TRIALS AND SEQUENTIAL TESTING

Charles R. Leake
USA Concepts Analysis Agency
ATTN: CSCA-RQR
8120 Woodmont Avenue
Bethesda, Maryland 20814

Abstract. A basic problem in conducting either clinical or sequential trials is to determine which or when statistical significance for a predetermined level of α has occurred. The criterion of

$$\alpha_{\tau} = \alpha/k$$

for k nonoverlapping comparisons is mentioned in a paper by Tukey (1). The consequences of not using this criterion are developed. The use of this criterion might be too stringent, however, and an alternative statistic is given.

Introduction. Tukey (1) presented a paper at the Birnbaum Memorial Symposium in May 1977. This paper was later published in Science. In this paper, Tukey mentions a criterion to determine whether or not one can say that he has observed statistical significance other than some random noise when making a number of comparisons on a set of data. This criterion, with all apologies to Professor Tukey, has been bestowed with the name Tukey's Criterion through common usage in a number of circles in the military analytical community.

The criterion is, for a given level of significance say α where k is the total number of plausible comparisons, $\alpha_{\tau} = \alpha/k$. Thus, if one observes a difference which has a probability of occurring of α_{τ} or less when one is comparing k nonoverlapping classes (or subsets) of a sample space, then one can say that this difference is statistically significant at the α level.

The converse shows why this is necessary. Table 1 gives a sample of the probability of not reaching statistical significance at $\alpha = .05$ (5%) and $\alpha_{\tau} = .05/k$ for a selected number of comparisons, as well as the probability of observing at least one statistical significance for $\alpha = .05$. Clearly for a fixed α level, the greater the number of comparisons which one makes, the more obvious it becomes that one will observe at least one statistical significance. Thus, the practice of conducting a test, making pair-wise comparisons, and reporting the significances for a fixed level

of α shows a certain statistical naivety. When done deliberately, it raises an obvious ethical question. To quote Tukey on this subject,

"The moral seems to me to be abundantly clear: Knowing that, for one class of patient, a clinical inquiry has reached some specific level of significance, such as 4%, is not evidence of the same strength as knowing that a focused clinical trial, involving a prechosen question, has reached exactly that level of significance, even if both the inquiry and the trial involved the same number of patients exposed to risk, and the same total number of end points, distributed in the same way." (1, p. 681)

Table 1. Sample of Probabilities of Not Reaching Significance

Sample number of comparisons	Probability of not reaching significance		Probability of at least one significance at 5%
	At 5%	At 5%/k	
1	95.0	95.0	5.0
2	90.2	95.1	9.8
3	85.7	95.1	14.3
4	81.5	95.1	18.5
5	77.4	95.1	22.6
10	60.0	95.1	40.0
20	35.8	95.1	64.2
50	7.7	95.1	92.3
100	0.6	95.1	99.4

What then can one do, when one is conducting an inquiry on a set of data that might not even have been created by the inquirer? There is one obvious answer to this question, use Tukey's Criterion to determine which comparisons are statistically significant.

In order to use Tukey's Criterion, one must first divide α by the number of comparisons to be made. Let's assume for illustrative purposes that $\alpha = .05$ and k , the number of comparisons is 20. It follows, then, that the α -level, adjusted for Tukey's Criterion becomes $\alpha_{\tau} = \alpha/k = .05/20 = .0025$.

Thus, the probability of rejecting H_0 is not .05 but .0025 when α is ad-

justed in accordance with Tukey's Criterion. The effect of this change in α -level is reflected by a corresponding change in the rejection region of the statistic being used. For example, if a Z-score is being used, for $\alpha =$

.05, the critical Z is 1.64. On the other hand, if $\alpha_T = .0025$, as Tukey's Criterion specifies, then the critical Z is 2.81. Thus, the observed difference must be over 1.1σ greater than would be required if the α were not adjusted for Tukey's Criterion. As a result, the data may not be compatible with such a requirement for statistical significance. Another would be to use another statistic such as Scheffe comparisons in conjunction with an analysis of variance. However, in order to use analysis of variance, there are certain data requirements such as equal or proportional cell size in a two or more way analysis of variance. That available data does not always lend itself to such an analysis goes without saying.

It appears more likely that choosing either of these alternatives is unsatisfactory to the inquirer. Either Tukey's Criterion is too stringent, or one does not have the required prerequisites for an analysis of variance or a similar nonparametric substitute. What then?

Alternative Statistic. An examination of the problem raised by Tukey leads to an alternative approach to attempt to attach meaning to making comparisons on a set of data. Consider the following problem:

How many observed statistical significances made on k, nonoverlapping, and statistically independent comparisons must be made in order to say that the number observed has less than a 5% probability of occurring?

The answer to this question can be found by using the binomial distribution and solving for x, where

$$b(x:N, 1-\alpha) < .05.$$

As shown in this inequality, x is the desired number of statistical significances, N the number of comparisons, and α , the significance level.

If this number of statistical significances is achieved, one could imply that factors other than chance were involved in obtaining that number of statistical significances. Moreover, this statistic could be used for parametric and nonparametric comparisons as well as a substitute method for an analysis of variance where such an analysis was unfeasible due to sample considerations.

Table 2, which was obtained by using the binomial theorem for $n \leq 100$, is shown below for $\alpha = .05$. For $n > 100$, a normal approximation of the binomial theorem can be used. The number of observed significances were obtained from a binomial table (2). This table, or the one below, can be used for $N \leq 100$ to determine whether or not the number of observed statistical significances occurred by chance alone.

Table 2. Number of Observed Statistical Significances for $\alpha = .05$ for N Comparisons to Occur with Less than 5% Probability

N, number of comparisons	Observed significances
1	1
2-7	2
8-17	3
18-28	4
29-40	5
41-53	6
54-66	7
67-79	8
80-96	9
97-100	10
n 100	Use normal approximation

REFERENCES

1. Tukey, J.W., "Some Thoughts on Clinical Trials, Especially Problems of Mutiplicity", Science, 18 Nov 77, pp. 679-684.
2. USAAMSAA, Cumulative Binomial Probability Distribution (AD 502418), 1979.

FIRE SUPPORT TEAM EXPERIMENT

**Jock O. Grynovicki
Jill H. Smith
Virginia A. Kaste
Ann E. McKaig**

**Experimental Design and Analysis Branch/ACE Team
Systems Engineering and Concepts Analysis Division
Ballistic Research Laboratory
Armament Research and Development Center/USAAMCCOM
Aberdeen Proving Ground, MD. 21005**

ABSTRACT

The Army is fielding a new digital communications system, the TACFIRE system, shown for the brigade-area in Figure 1. In order to investigate the command, control, and communications issues associated with the new devices, the Artillery Control Environment (ACE) was developed. ACE is a real-time, multiplayer, interactive simulation system run on a commercial computer that interfaces with the tactical equipment through a bit box (modem). This paper discusses the preparations, experimental design, data collection, analysis methods, and results for the first experiment with military players interfaced with the Artillery Control Environment software conducted 8 May - 10 June 83.

BRIGADE-AREA TACFIRE SYSTEM

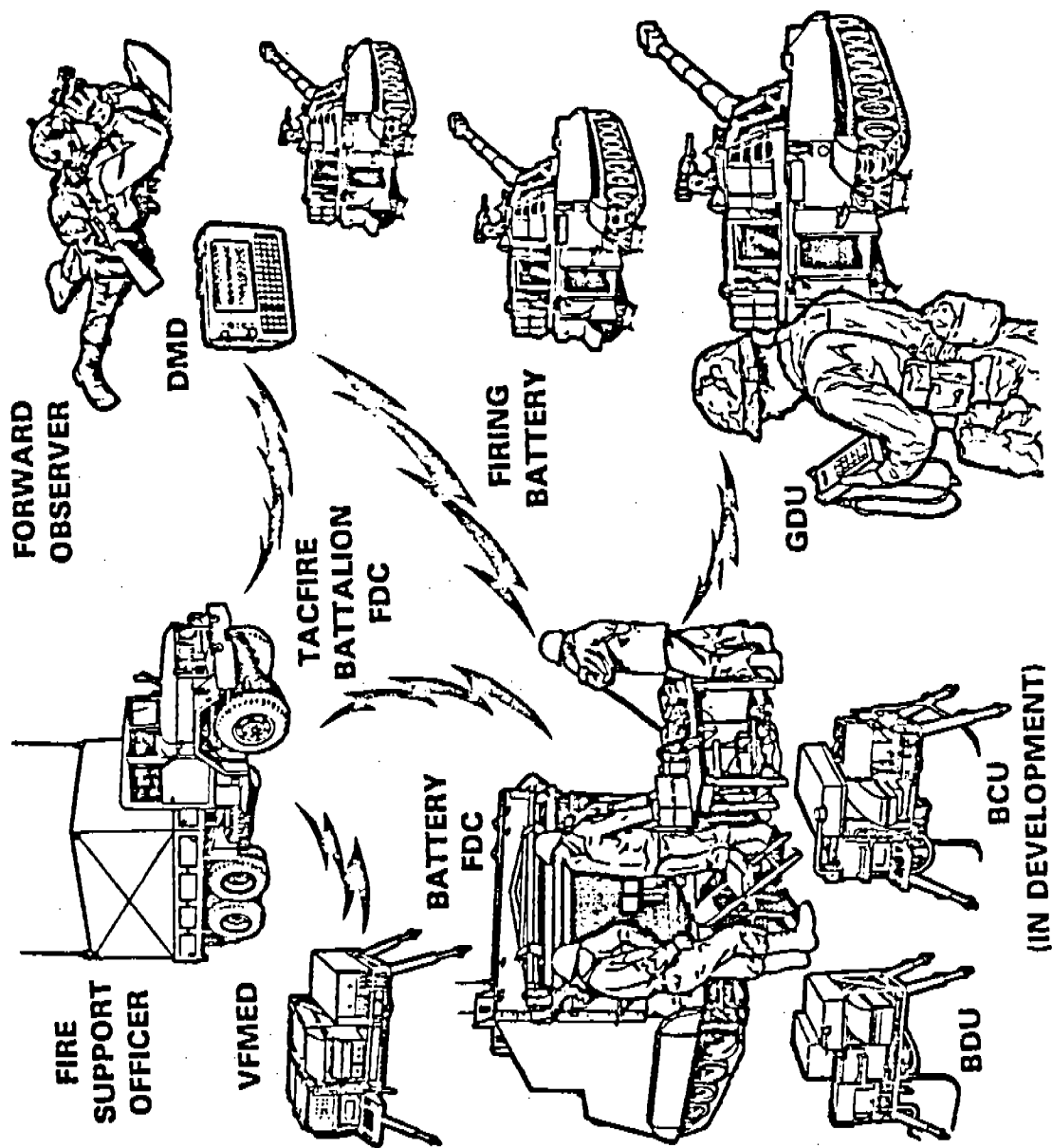


FIGURE 1

I. INTRODUCTION

A. Background

In May 1982, the HELBAT (Human Engineering Laboratory Battalion Artillery Test) Executive Committee agreed that the Ballistic Research Laboratory Artillery Control Environment (ACE) and HELBAT activities should be combined to develop a Command Post Exercise Research Facility (CPXRF). The CPXRF will primarily be used for research, development, testing and evaluation (RDT&E) work in automatic data processing (ADP) fire support control systems using commercial ADP technology; a secondary usage is the training of the tactical ADP operators under controlled conditions. Further, an ACE/CPXRF Subcommittee was formed to provide joint DARCOM-TRADOC guidance in the development of ACE technology and use of the CPXRF. The ACE software is a key tool in the CPXRF. The software features the ability to automatically load live players with messages produced by target acquisition and fire direction simulators while recording all the message traffic that flows between the live and simulated players.

An overview of the CPX Research Facility and ACE program is given in the 1982 Sept-Oct issue of the Field Artillery Journal in an article "HELBAT/ACE Fire Support Control Research Facility" by Mr. Barry Reichard. The layout of the facility is shown in Figure 2.

B. Purpose

The experiment detailed in this report was the first test in which military players were interfaced with the Artillery Control Environment (ACE) software. The purpose of this experiment was to demonstrate the feasibility of using the automated techniques of the CPX Research Facility for fire support control experiments.

To demonstrate this capability, a study of the effects of message intensity and communication degradation on the Fire Support Team Headquarters' (FIST HQ) ability to perform fire support coordination was performed. Message intensity was defined to be a function of message type, message rate, and message content.

II. TEST CONCEPT

A. Objectives

1) To determine the effect of message intensity on the FIST HQ's ability to perform fire support coordination.

2) To determine the effect of communication degradation on the FIST HQ's ability to perform fire support coordination.



COMMAND POST EXERCISE RESEARCH FACILITY

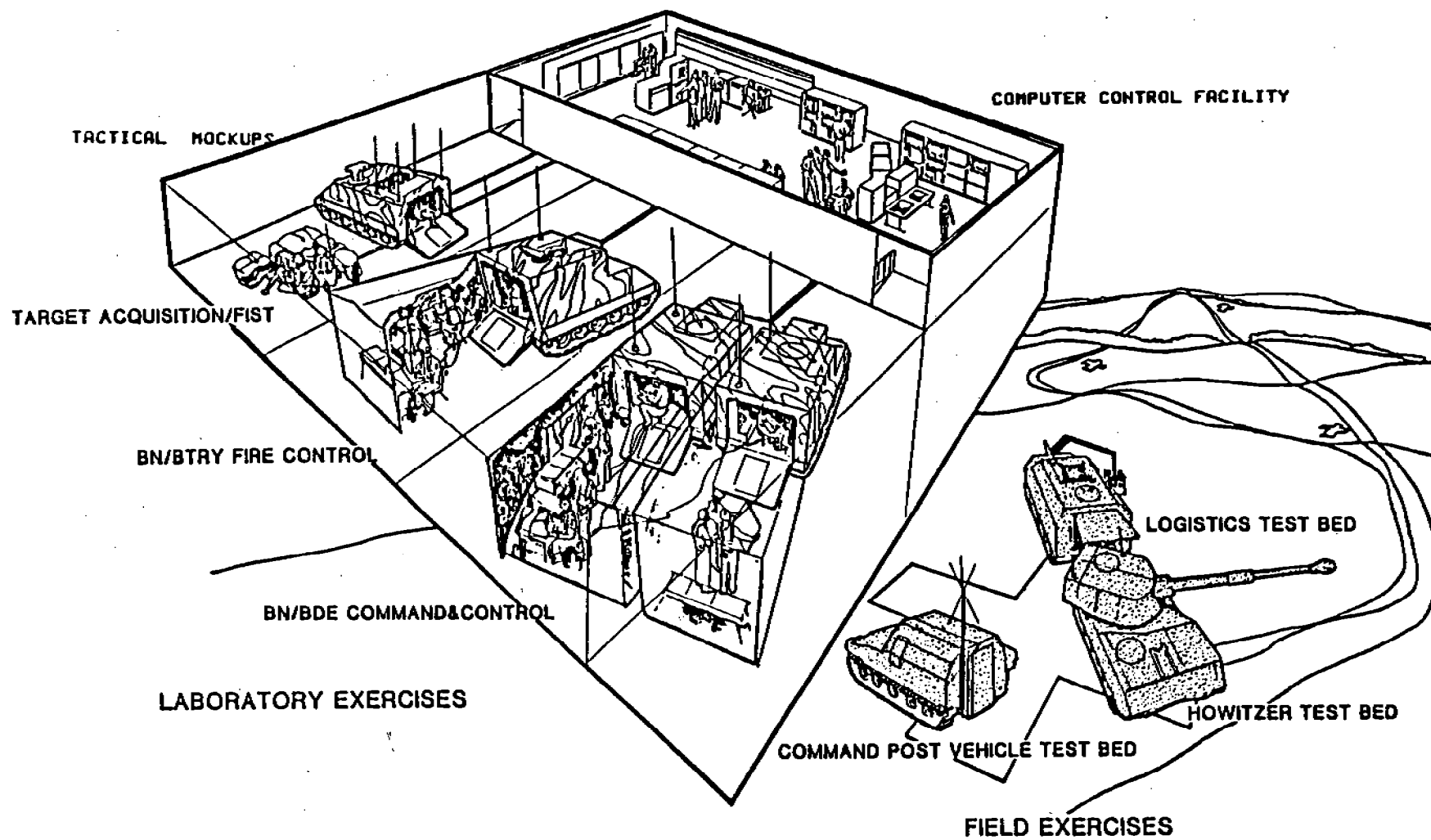
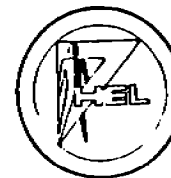


FIGURE 2

3) To determine if message intensity and degraded communication have a combined effect on fire support coordination.

B. Measures of Performance

A measure of performance (MOP) is a response that is used to quantify the effects of the factors to be evaluated. Because all of our objectives investigated the effect on fire support coordination, the measures of performance were the same for all three objectives. The following measures of performance were computed on each two hour cell of the test:

1) Number of messages serviced by the FIST HQs (i.e. messages for which a response was generated). This number provides information on the message traffic at the FIST HQs under the different conditions and can be translated into net usage.

2) Service time distribution, where service time is defined to be the time required for the FIST HQ to service a message starting from the time the ACK is sent from the FIST DMD acknowledging receipt of a message to the time the response message is first transmitted. This measure indicates the combined time a message spends in the FIST DMD message queue and the processing and decision time of the FIST HQs.

3) Manual transmission time distribution, where manual transmission time is defined to be the time from first transmission of the response message by the operator to the time an acknowledgement (ACK) is received for that message. The FIST HQs have completed the decision making at this point, but must continue to send the message until an acknowledgement is received. In degraded communications this time may not be inconsequential. Also, the FIST HQs cannot process other messages while transmitting manually.

4) Frequency count by number of tries for messages acknowledged. The FIST DMD has a one character field for try number that cycles modulo 4 (i.e. 0,1,2,3,0,1,2,3,0,...). It was noticed in HELBAT 8 data, that more than four tries were sometimes necessary to get an acknowledgement back on a message. TACFIRE uses the try number in the FIST DMD message to determine what authenticator to select for comparison to the DMD message. Therefore, if the number of tries exceeds four the FIST DMD displays a message to the operator to contact destination by voice to resynchronize the authenticator codes. This voice-digital contention then causes more problems to a net that is already experiencing communications problems.

5) Number of fire missions completed/number of fire missions initiated. The FIST HQs were given two hours and ten minutes to complete two hours of scenario. A completed fire mission, by definition, is a call for fire (FR GRID), a message-to-observer (MTO), at least one SHOT and an end-of-mission (EOM).

6) Number of fire missions completed/number of fire missions expected. The number of fire missions expected is the number of fire missions in the database. This was to measure if the FIST HQs could complete all fire missions in the two hour scenario database within the two hours and ten minutes allotted.

C. Scope

The fire support team was a four-man team consisting of:

- 1) the fire support team chief
- 2) the fire support sergeant
- 3) two radio telephone operator/drivers.

The FIST chief was available to the FIST HQ for initial supervision only. As per typical operating procedures, the FIST chief may be absent for extended periods of time (hypothetically accompanying the company commander).

The FIST HQ was task-loaded by software interactively simulating three platoon-level forward observers. The software FOSCE (Forward Observer SCENARIO) used tactical scenarios developed by Mr. Arthur Long of the US Army Field Artillery Board. This scenario or input database is detailed in the Section III-D, "Input Data Base".

The FIST HQ had direct access to fire support from a company-level mortar platoon fire direction center (FDC) and a generic field artillery fire direction center. All FDC operations were simulated interactively by software. The FIST HQ determined the proper action (based on the FIST chief's guidance and training) for each fire request; either to deny the request, service the request with mortars or forward the request. Fire support was unlimited, that is, not constrained by ammunition resupply.

All members of the FIST Headquarters were trained in the operation of the FIST DMD to give the FIST chief flexibility in managing his team.

D. Limitations

- 1) All observers were placed in the review mode in the FIST DMD subscriber table.

2) After deciding a fire request should be handled either by the mortars or forwarded to the FDC, the fire mission was forwarded in the automatic mission mode. That is, all subsequent messages for that fire mission are automatically routed through the FIST DMD. Operator intervention is needed only if a message does not get acknowledged in four tries. He is then notified that a message did not get ACKed after four tries, the message is placed in his message queue and must be forwarded manually.

3) No FIST HQ initiated missions were included.

4) No tactical chores were performed, e.g., guard duty, close station march order, emplacement, etc.

5) All communication was digital, no voice communication.

E. Test Configuration

Figure 3 shows the nodes that were played in the first military player test. The FIST HQ equipped with the FIST DMD in the mock-up vehicle interacting through ETHER, the intracomputer communications network, with three forward observer scenario programs, the mortar fire-direction simulator and battalion fire-direction simulator. Figure 4 shows how these players communicated together and the net assignments.

III. RESOURCE REQUIREMENTS

A. Software

ACE software permits real-time fire support command and control functions to be exercised in a controlled laboratory environment. The software is written in the C programming language and is designed to run under the 4.1bsd (Berkeley) UNIX operating system. The major components of the ACE software are described below.

1. ETHER ETHER is a single program which functions as an intra-computer communications network. Computer ports are assigned to communication nets. ETHER accepts a message from a port and transmits it to all other ports on the assigned net. Message collisions are prevented by separately buffering each message within ETHER.

Each net is assigned a probability of message loss which ranges from zero to one. If the probability of message loss is zero, the net was an ideal net and all messages are sent to each port on the net. If the probability of message loss is greater than zero, a uniform random number generator is used to decide whether or not a message is lost. Lost messages are not transmitted to any port on the net. Acknowledgements are treated the same as any other message.

ETHER maintains a log file of each message which it receives. In addition to the raw message, the log contains the times (Julian day, hour, minute, second) for the start of the message, the end of the preamble and the end of the message.

2. Ace Display (ADIS). ADIS utilizes a CRT (cathode ray tube) terminal to display in real time the messages being transmitted through ETHER. The terminal screen is divided into eight columns which are labeled for the players (see Figure 5). Each message is displayed as two lines in both the sender's and receiver's columns. The message first appears in the sender's column. The first line contains the message type and target number if it has

FIST HEADQUARTER EXPERIMENT

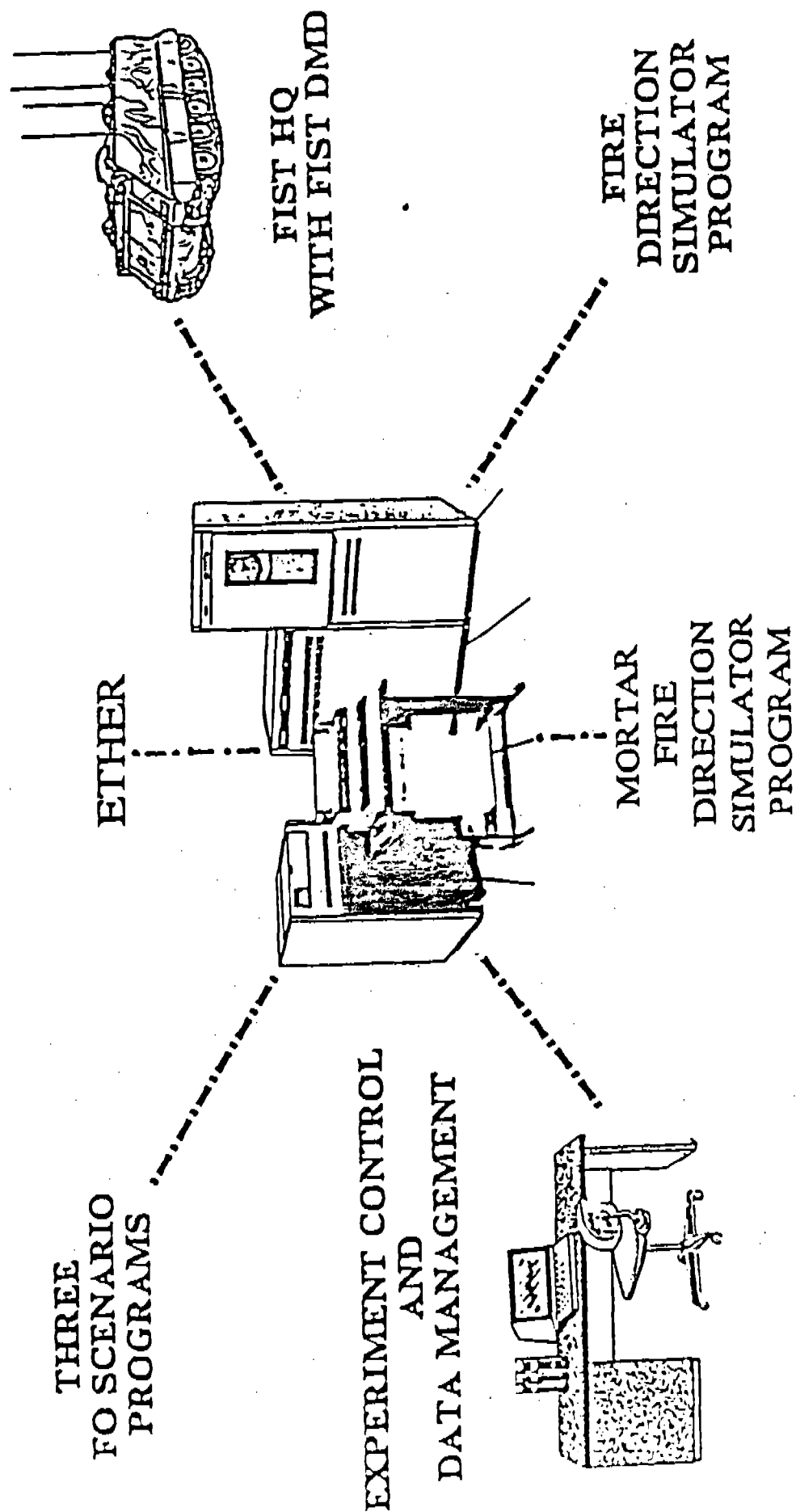


FIGURE 3

ETHER: Intra-computer communications network
 FOSCE: Forward Observer SCEnario program
 FDS: Fire Direction Simulator program
 MFDS: Mortar Fire Direction Simulator program

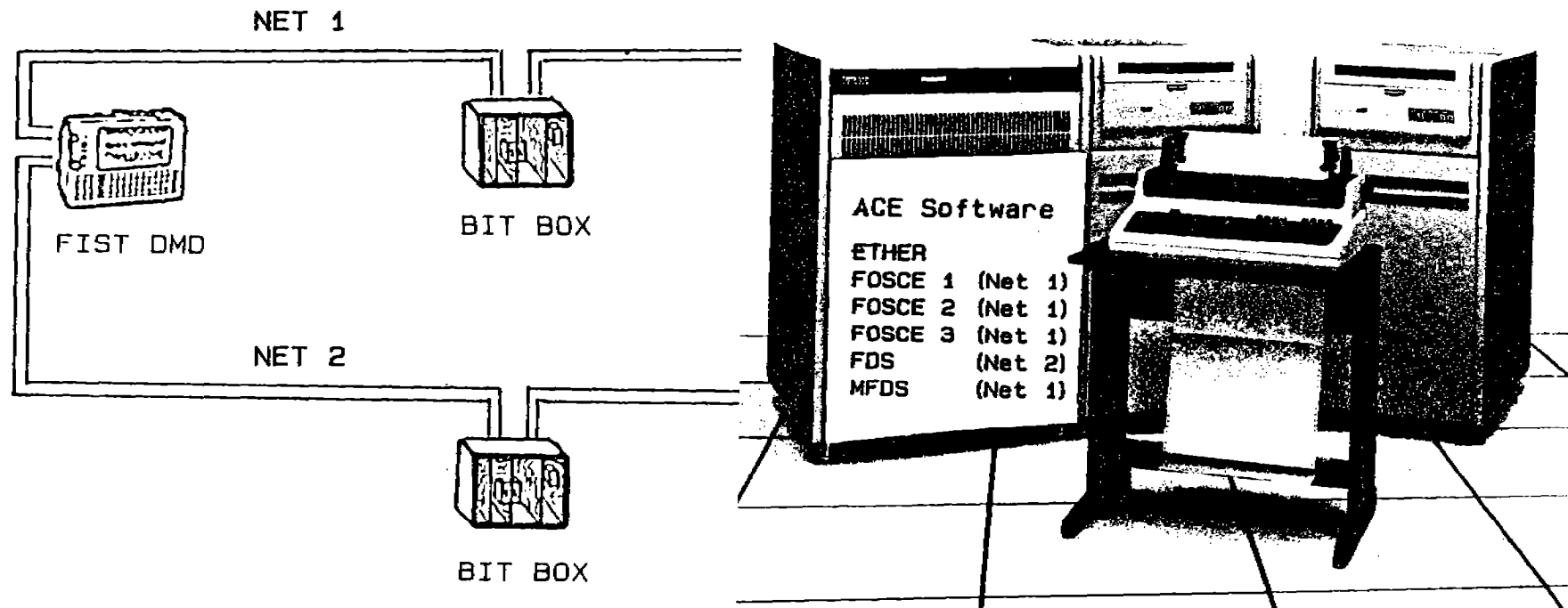


FIGURE 4. Test Configuration

FO	1IFO	2IFO	3IFIST	FIFDS	VIMFDS	BI	al
FR GRID			IFR GRID				
*^ 10:19			*1 10:21				
			IFR GRID		IFR GRID		
			*^ 11:06		*F 11:08		
			IMO AF3700		IMO AF3700		
			*B 11:43		*^ 11:41		
MSG LOST			IMO AF3700				
F			^ 11:45				
MO AF3700			IMO AF3700				
*F 11:54			*^ 11:52				

*****retry 1*****

transmitter -> receiver : - - - F -> 1
message type : - - - MTO
target number : - - - AF3700
transmit /107:00:11:52\ end preamble /107:00:11:53\ end msg /107:00:11:53\
msg : 10AADFS10603700 30 1 10401000111 008400100

/000:00:11:54\

FIGURE 5. Sample ACE Display (ADIS)

been assigned. The second character in the second line is a "", indicating "sender" and the time sent is given. The message will then appear in the "receiver's" column. The first line is the same as in the "sender's", the second character in the second line gives the address of the "sender" and the time received is displayed. When the acknowledgement is sent by the "receiver" an "" is displayed as the first character in the second line of the "receiver" and when the acknowledgement is received by the "sender" an "" is displayed as the first character in the second line of the "sender". If the message is degraded by ETHER "MSG LOST" appears in the receiver's column. Below the columns, the last message sent is interpreted. At the bottom of the screen, the time from the start of that run is displayed.

3. Forward Observer Scenario (FOSCE). Forward observer scenario program reads a database of forward observer (FO) messages and transmits the messages as if they were being generated by a real FO. Each message is time-tagged in the database and sent by FOSCE at the appropriate time. FOSCE will retransmit a message up to four times if an acknowledgement is not received. FOSCE, after sending a request for fire, will wait for a message-to-observer (MTO) and SHOT message before transmitting subsequent adjust (SA) messages. Because no voice communication was allowed, FOSCE was made smart enough to respond to freetext messages asking for the status of a particular fire mission by target number or the status of FOSCE itself, that is, active or not active.

4. Fire Direction Simulator (FDS). The fire direction simulator consists of four programs which perform a limited number of TACFIRE/BCS functions. FDS accepts fire request messages, prioritizes them, assigns target numbers and generates MTO and SHOT messages. The number of simultaneous missions which the FDS will process may be specified. If the number of missions exceeds the maximum, the FDS will process missions based on mission priority. During this experiment, the FDS could handle up to 10 missions simultaneously, which was not a limitation on the system. The FDS could be queried by the FIST HQs as to the status of a particular fire mission by target number or by observer identification number and mission buffer.

5. Mortar Fire Direction Simulator (MFDS). The mortar FDS simulates communication with the 81 mm company mortars. It is a special version of the FDS program which will only accept one fire mission at a time.

6. Bit Box Program (BBP). The bit box interface program accepts messages from ETHER and transmits them to a computer port which is connected to a bit box. The program also reads messages from the computer port and transmits them to ETHER.

B. Hardware

1. Two Bit Boxes. Bit boxes are microprocessor based modems which enable TACFIRE hardware to communicate with commercial computers. Bit boxes accept TACFIRE messages from wire line or radio, perform error correction and convert the messages to RS232 ASCII characters which commercial computers can accept. They will also accept a message from the computer, add the error correction bits, time disperse the message and transmit it over wire line or radio in TACFIRE format (FSK).

2. FIST DMD. The FIST digital message device that was used in the experiment is one of four experimental design models (EDM #2) that are in existence. It is a prototype model, and not a production model.

3. VAX 11/750 Computer. The VAX 11/750 computer was dedicated to running the experiment and had no other processes running during the test. The operating system was the 4.1bsd (Berkley) UNIX.

C. Training

Test participants were collectively trained at the Human Engineering Laboratory in the operation of the FIST DMD by CPT Gahagan, an instructor from the Gunnery Department of the US Army Field Artillery School. The Human Engineering Laboratory provided training equipment for the students. The test participants were trained Fire Support Teams (MOS 13F) from the 82nd Airborne Division, Ft. Bragg.

D. Input Database

The tactical scenario database contained all fire support control messages for a limited scenario of a mechanized infantry battalion of an armored division. The SCORES, Europe III, Sequence 2A was used to generate targets expected to be fired by a battalion in sustained combat operation. The battalion is constrained by ammunition resupply under normal operations, however, it was decided that ammunition resupply should not be a limiting condition in this test. The entire scenario was played in retrograde mode.

The data base consisted of 36 two hour cells of messages, 12 two hour cells of low intensity, 12 two hour cells of medium intensity and 12 two hour cells of high intensity. Intensity is defined by the number of initiating messages per two hour cell as given in Figure 6 and the message stream that follows each initiating message as given in Figure 7. It can be seen that intensity is a function of the number of initiating messages and their subsequent messages. The 36 two hour cells of data were arranged such that all permutations of the three intensities (L-M-H) appeared twice. Ninety percent of the fire missions had normal priority and the other ten percent had urgent priority.

IV. DATA COLLECTION

A. Experimental Design

1. Factors The two factors that were tested in this experiment are message intensity and communication degradation. Three levels of message intensity were tested with each of the three levels of communication degradation giving nine test combinations. The levels of each factor were defined as follows:

FACTORS

1) INTENSITY (per two hour block)

<u>MESSAGE TYPE</u>	<u>LEVELS</u>		
	Low	Medium	High
Fire Mission 1, Fire For Effect	4	8	12
Fire Mission 2, Adjust Fire	2	4	6
Fire Mission 3, Immediate Smoke	0	1	1
Artillery Target Intelligence	18	12	6

2) COMMUNICATION DEGRADATION

00% Message Loss
15% Message Loss
30% Message Loss

FIGURE 6

INTENSITY			
MESSAGE SEQUENCE .	LEVELS		
	L	M	H
1) Artillery Target Intelligence ATI FO → FIST → FDC	18	12	6
2) Fire Mission, Fire for Effect: FR GRID FO → FIST → FDS MTO FO ← FIST ← FDS SHOT FO ← FIST ← FDS EOM FO → FIST → FDS	4	8	12
3) Fire Mission, Adjust Fire FR GRID FO → FIST → FDS MTO FO ← FIST ← FDS SHOT FO ← FIST ← FDS SA(1) FO → FIST → FDS SHOT FO ← FIST ← FDS SA(2) FO → FIST → FDS SHOT FO ← FIST ← FDS SA(3) FO → FIST → FDS SHOT FO ← FIST ← FDS EOM FO → FIST → FDS	2	4	6
4) Fire Mission, Immed. Smoke Same as Adjust Fire Mission	0	1	1

FIGURE 7

Message Intensity

L = low

M = medium

H = high

Communications Degradation

0 = 0% degradation

1 = 15% degradation

2 = 30% degradation

2. Design Matrix It was decided that the smallest period of time reasonable to test any one of the nine treatment combinations was two hours. Since the testing of all nine treatment combinations require a minimum of 18 hours of testing and realistically could not be completed in one day, a randomized incomplete block design was constructed so that the day-to-day variability would not influence the results. The nine treatment combinations were divided into blocks of three and the three blocks were run over a three day period. The assignment of the treatment combinations into blocks was based on a confounding scheme. This scheme assures that the effects of message intensity (I) and communication degradation (C) and the interaction of these two factors (I x C) on a FIST HQ's ability to perform fire-support coordination can be measured. Because time constraints permitted only two replications, part of the precision of the estimate of the interaction was sacrificed (i.e. blocks within replicate 1 were confounded with the linear component of the I x C interaction and blocks within replicate 2 were confounded with the quadratic component of the I x C interaction). Randomization of treatment combinations within blocks and blocks within days was performed.

The experiment was repeated for four FIST teams, so that team-to-team variability was included. In addition, software changes were implemented between teams 2 and 3 as a result of information from a pilot test. The pilot test was conducted before the actual test and resulted in changes to the software to make it tactically more realistic. One significant change was to have the FDS send one SHOT message per call-for-fire rather than one SHOT message per volley. Capability for status requests was implemented in the FDS at this time also. Because of these changes, software was made a factor in the experiment so that the variability could be accounted for due to the software changes.

The design matrix is shown in Figure 8. The FIST teams were tested sequentially, one at a time for six days. The six days are shown in the design matrix and the tests were run in the order given within each day.

DESIGN MATRIX							
SOFTWARE	FIST	REP1			REP2		
	TEAM	DAY1	DAY2	DAY3	DAY4	DAY5	DAY6
S1	TEAM ONE	L2 M1 H0	M0 H2 L1	L0 H1 M2	M0 H1 L2	L0 M1 H2	H0 L1 M2
	TEAM TWO	H1 L0 M2	H2 M0 L1	L2 H0 M1	M1 L0 H2	M2 H0 L1	H1 M0 L2
S2	TEAM THREE	M2 H1 L0	H2 M0 L1	M1 L2 H0	L0 M1 H2	H1 M0 L2	M2 H0 L1
	TEAM FOUR	H2 M0 L1	M1 L2 H0	M2 H1 L0	H1 M0 L2	L1 H0 M2	L0 M1 H2

INTENSITY

L= LOW
M= MEDIUM
H= HIGH

COMMUNICATION DEGRADATION

0= 00% DEGRADATION
1= 15% DEGRADATION
2= 30% DEGRADATION

Figure 8

V. DATA ANALYSIS

A. Statistical Analysis

1. Effect of Factors on Message Traffic. The total number of messages generated for each experimental condition over a two hour cell was used to evaluate and validate the effect that the different factors and their interactions had on message traffic. Based on the way intensity and communication degradation were defined in planning this experiment, we would expect these two factors to have a significant effect on message traffic. An increase in intensity levels should result in an increase in the number of messages generated. Similarly, an increase in communication degradation should result in an increase in the number of messages it takes to complete a fire mission or to forward an artillery target intelligence message. To some this may seem counter intuitive, however, in degraded communications the messages are being sent but not received and this results in retransmissions increasing the message traffic. The other factors specified in the design, including the two different Fire Direction Simulator software programs, were also included in this analysis.

The number of messages observed in each test cell are shown in Figure 9. An analysis of variance was performed on this measure with all replicate interaction terms pooled for the error term. A second analysis of variance procedure was then performed with additional interaction terms found not to be significant also being pooled with error. The ANOVA table for the final reduced model is shown in Figure 10. It should be noted that since block was confounded with components of the intensity-degradation interaction, it is not meaningful to test any term in the model containing block. A star next to the F-statistic indicates that the factor is significant. Based on the calculated F-values, intensity, degradation, intensity-degradation interaction, software, and intensity-software interaction, were found to have a significant effect on the message traffic.

The effect that intensity, degradation and their interaction have on message traffic is summarized in Table 1. Table 1 gives the average number of messages per two hour cell, μ , and the number of cells in the average, N, for the given factors and their marginal effects (averages over the rows and columns). Looking at the average number of messages generated for each level of intensity presented in the right hand column of Table 1, one sees that there is a significant increase from 361.46 to 882.50 as intensity increases. Similarly, an increase in communication degradation increased the average message traffic flow from 462.13 to 798.58. In addition, in comparing the mean change between the different levels of communication degradation for each level of intensity, a positive interaction effect can be noted. There was an increase in the mean of about 200 messages between 00% and 30% degradation for low intensity compared to an increase of over 300 messages for medium and 500 messages for high intensity.

The effect that software and the software-intensity interaction has on message traffic is summarized in Table 2. The average number of messages generated per two hour block for the original FDS software program was 704.67 compared to 545.83 for the modified program. The software was changed to produce a shot message every call for fire instead of every

Software	Intensity	Fist Team	Communication Degradation						Total
			00		15		30		
			RepI	RepII	RepI	RepII	RepI	RepII	
S1	L	Team1	297	281	410	335	508	494	4648
		Team2	285	279	353	413	516	475	
	M	Team1	512	468	742	723	761	782	8531
		Team2	552	564	781	649	1146	852	
	H	Team1	811	700	890	1191	1276	1303	12368
		Team2	778	808	996	1076	1216	1323	
S2	L	Team3	300	230	329	288	508	441	4097
		Team4	245	238	316	390	393	419	
	M	Team3	393	396	599	512	727	750	6722
		Team4	411	402	556	558	730	688	
	H	Team3	530	544	624	746	1005	972	9004
		Team4	563	548	726	701	881	1104	
TOTAL			11135		14966		19270		45370

Figure 9

ANALYSIS OF VARIANCE (EFFECT ON MESSAGE TRAFFIC)

ANALYSIS OF VARIANCE (ANOVA)				
SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	F RATIO
Replication	1	288.00	288.00	0.08
Software	1	454104.50	454104.50	128.19**
Block within Rep	4	67391.34	16847.54	
Software X Block within Rep	4	8701.11	2175.28	
Team within Soft X Block within Rep	8	40628.52	5078.56	
Intensity	2	3259353.58	1629676.79	460.04**
Software X Intensity	2	152010.75	76005.37	21.46**
Degradation	2	1362202.08	681101.04	192.27**
Intensity X Degradation	4	103933.83	25983.41	7.33**
Pooled Error	43	152325.79	3542.46	
Total	71	5600939.55		

Figure 10

TABLE 1. Intensity by Degradation

**Average Number of Messages
Per Two Hour Cell**

Intensity	Communication Degradation (%)			
	00	15	30	
LOW	8 266.88	8 351.88	8 465.63	24 361.46
MEDIUM	8 461.00	8 636.00	8 798.38	24 631.79
HIGH	8 658.50	8 857.25	8 1131.75	24 882.50
	24 462.13	24 615.04	24 798.58	N μ

volley which is a more realistic representation of how TACFIRE/BCS functions. Therefore, one would expect the average message flow to be less for software 2 than 1. Also, one would expect a greater change between low, medium and high intensity for software 1 than 2. From Table 2, the difference between means for low and high intensity for software 1 is over 600 messages compared to a difference of less than 400 for the modified software. To obtain a realistic description of the effect that message intensity and communication degradation have on network message traffic flow and on the Fire Support Teams' ability to perform effective fire support coordination, the analysis from this stage on will be based on the second half of the experiment using the modified software.

TABLE 2. Software by Intensity

**Average Number of Messages
Per Two Hour Cell**

Software	Intensity			
	Low	Medium	High	
1	12 386.5	12 707.58	12 1019.02	36 704.67
2	12 336.42	12 556.00	12 745.08	36 545.83
	24 462.13	24 616.04	24 798.58	N μ

2. Frequency Count by Number of Tries of Messages Acknowledged. Theoretically, the number of tries it takes for a message to successfully reach its destination and for an acknowledgement to be received by the sender should only be affected by the percent of communication degradation in the communication networks. Providing one knows what the actual percent degradation is, one can determine the theoretical distribution of how many times a message is sent before it is acknowledged for each level of communication degradation. When there is no communication degradation, one would expect that all messages would be acknowledged on the first try. In 15% degradation the probability that a message gets through and is acknowledged on any try is $(1-.15)(1-.15) = .7225$. The probability that a message does not get acknowledged on a given try is $1-.7225$. Using these probabilities, we can compute the probability that a message is acknowledged in a given number of tries. Table 3 gives the theoretical distributions for the probability a message gets acknowledged in n tries for 15 and 30% degradation.

Using the theoretical probabilities from above and the total number of messages actually acknowledged under each degradation level, we can check the actual effect of communication degradation with the expected effect as a check on the laboratory system. Figure 11 shows the distribution of messages acknowledged by try number in "perfect" communications (0%

TABLE 3. Theoretical Distributions for Messages Acknowledged

Number of Tries	General Formula	Degradation Level	
		15%	30%
1	p	.7225	.4900
2	$p(1-p)$	.2005	.2499
3	$p(1-p)^2$	.0556	.1274
4	$p(1-p)^3$	.0154	.0650
.	.	.	.
.	.	.	.
.	.	.	.
n	$p(1-p)^n$	.	.

degradation) for software 2. "Perfect" communication was not quite perfect since the bit boxes did not have net monitoring and message collisions resulted. Figures 12 and 13 give the same distributions for 15 and 30 percent degradation. Very good agreement was observed, and as shown in Appendix B, when tested statistically the distribution of messages acknowledged by number of tries is a function of communication degradation only and is not influenced by intensity, team variability or learning.

3. Time Required to Service a Message. This section investigates the effect that degradation and intensity had on the time it took for FIST HQs 3 and 4 to service a fire request (FR) message and an artillery target intelligence (ATI) message. Since fire requests are given a higher priority than ATI's and require more processing by the FIST team, message type had to be considered a factor in this analysis.

As the data was being checked for completeness, it was noted that the distribution of service time was skewed and that the variance of the observations under various experimental conditions exhibited discrepancies. A check for homogeneity of variance using Bartlett's test confirmed the latter observation. In addition, several experimental groups had observations that were extremely large (over four standard deviations from the group mean) and atypical of the majority of the service times observed under the same experimental conditions. These

NUMBER OF
MESSAGES
(12 CELLS)

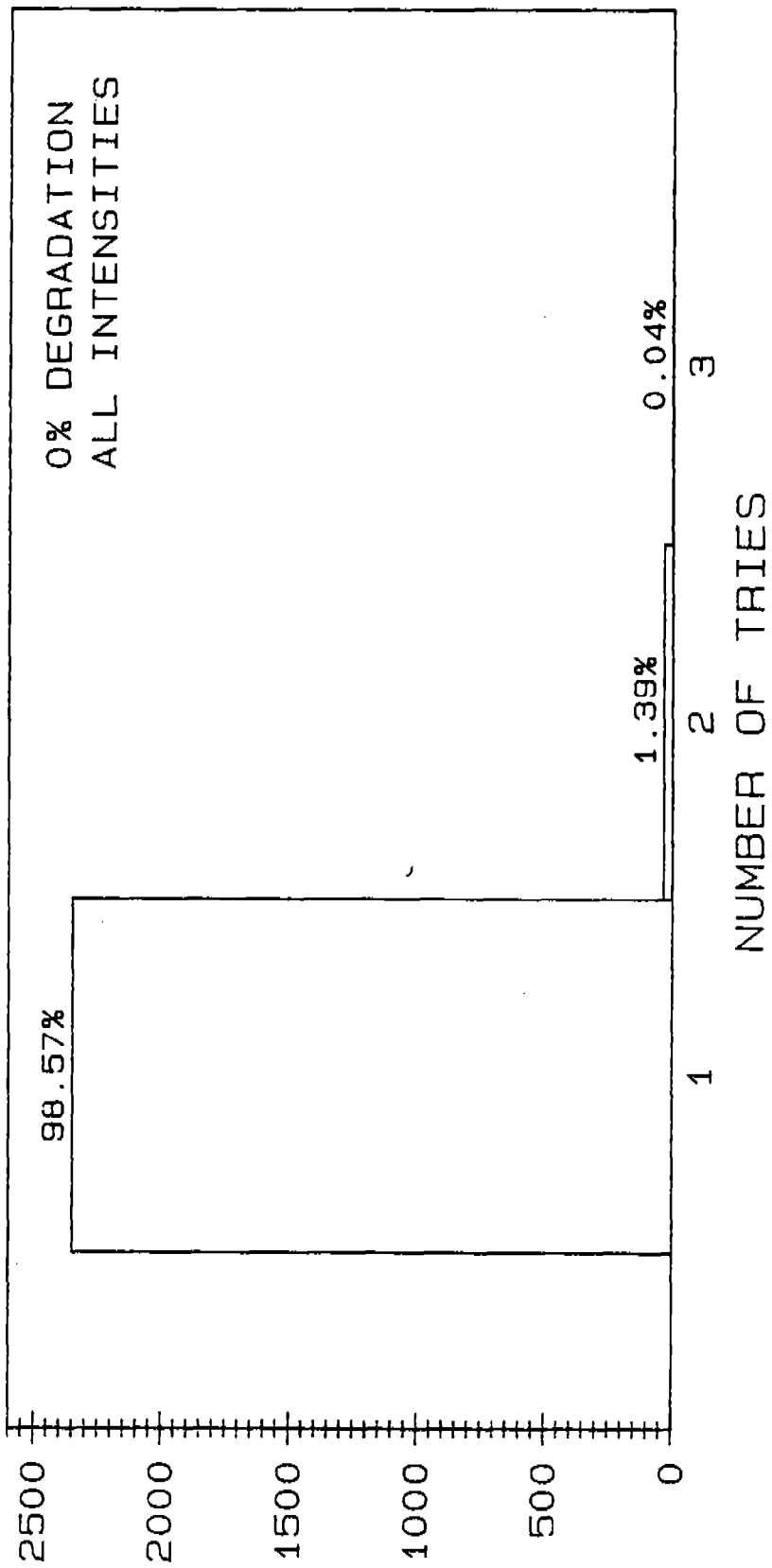


Figure 11

NUMBER OF
MESSAGES
(12 CELLS)

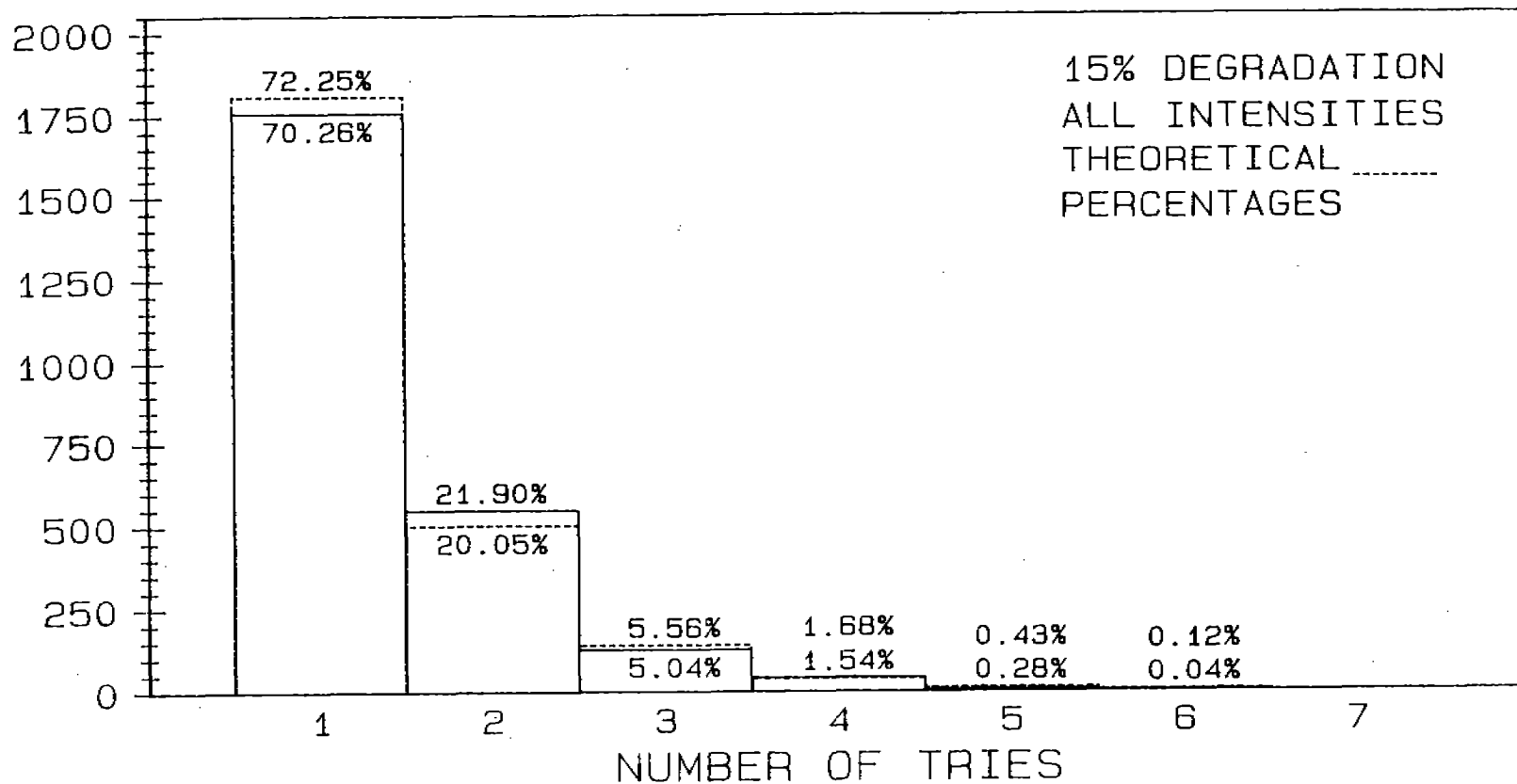


Figure 12

NUMBER OF
MESSAGES
(12 CELLS)

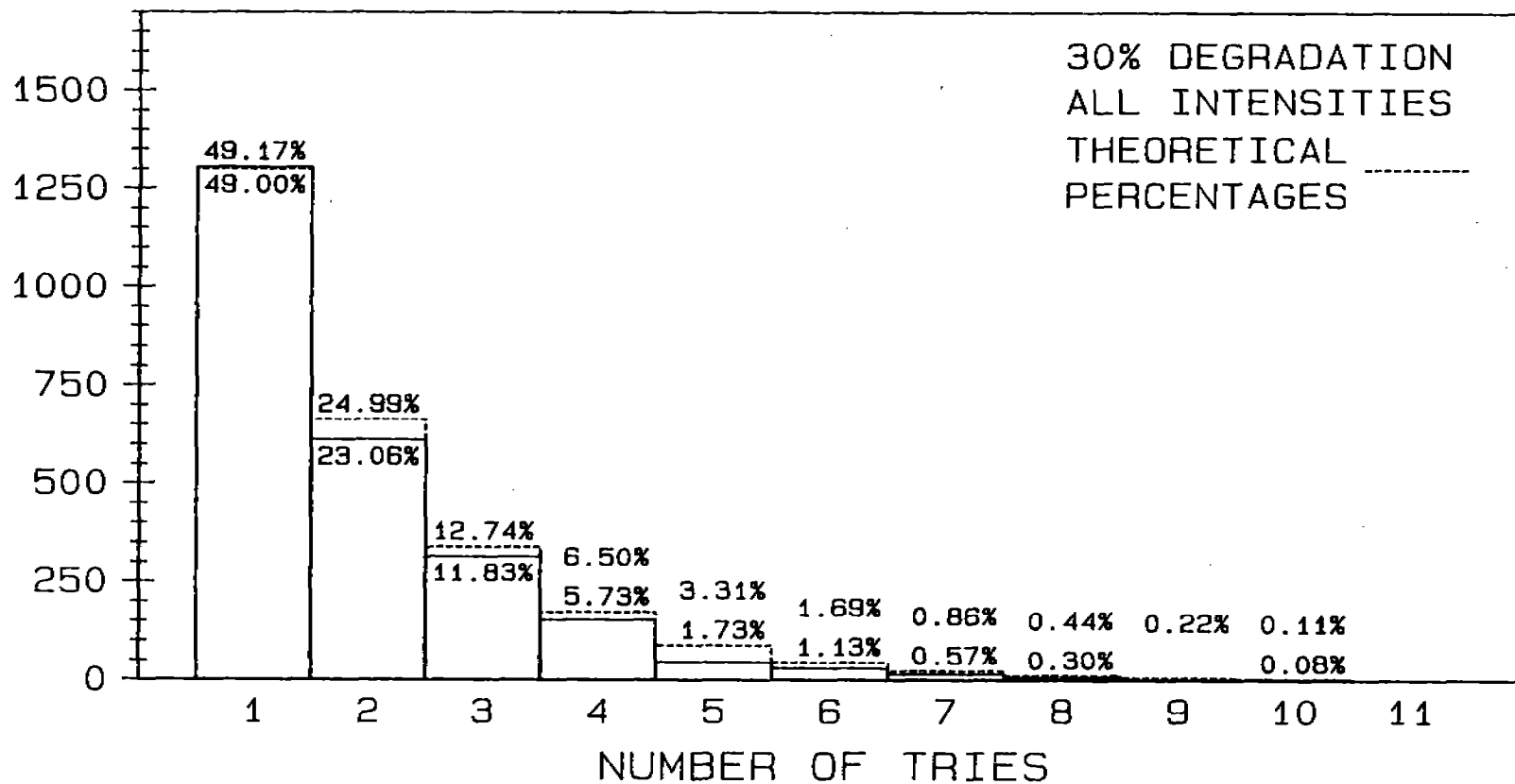


Figure 13

observations comprised slightly more than four percent of the total service times observed. They were removed from the analysis of variance procedure found below, but were considered in interpreting the final results. The median for each experimental condition with the outliers removed is given in Table 4 below.

TABLE 4. Median Service Time
by Experimental Condition

Rep	Message	Team	Intensity	Degradation		
				0	15	30
1	ATI	3	L	9.2	12.0	27.0
			M	10.5	14.0	8.5
			H	9.0	14.5	23.0
		4	L	9.3	6.1	6.0
			M	6.5	6.5	9.0
			H	6.5	9.0	9.0
2	ATI	3	L	9.0	4.2	9.2
			M	9.5	10.3	9.5
			H	3.5	8.3	40.0
		4	L	6.3	7.8	5.5
			M	5.3	9.0	8.1
			H	7.5	6.5	11.5
1	Fire Request	3	L	15.5	22.0	46.0
			M	18.3	20.5	15.0
			H	17.3	16.0	21.5
		4	L	12.5	8.0	9.0
			M	6.3	8.5	9.3
			H	6.7	11.0	10.5
2	Fire Request	3	L	14.5	14.5	18.3
			M	14.3	16.7	18.5
			H	13.3	14.5	22.8
		4	L	8.0	9.5	8.0
			M	9.8	10.9	8.4
			H	11.3	8.8	17.5

Further investigation of the data revealed a positive correlation between the standard deviations and the experimental group means. Correlation between the standard deviations and group means is often accompanied by marked non-normality and non-homogeneity of variance, and indicates that the particular form of the original observations is unsuitable for ANOVA procedures. However, a transformation can be determined which makes the standard deviation independent of the mean, corrects non-homogeneity and also results in the observations being distributed more normally. In general, if a significant functional relationship between the standard deviation and the group means can be determined, then the transformation is the integral of the reciprocal of this functional relationship. Following this procedure, the following transformation was developed:

$$1.3 \ln(-2.6 + .8(\text{service time}))$$

The transformed data became more normal and the assumption of homogeneity of variance was confirmed.

An analysis of variance procedure was performed on the transformed data. One slight modification to this procedure was that due to unequal experimental group sizes, the sum of squares for all terms in the model, except the error term, was weighted by the harmonic mean. The final reduced ANOVA Table is presented in Table 5.

The most significant term in the analysis was team. The median service time for team 3 was 14.5 seconds which is substantially higher (73 percent) when compared to the 8.5 seconds for team 4. This trend is prevalent for both fire requests and ATI messages, but is magnified when one considers just fire requests. As suspected, type of message also influenced service time. Although fire requests have a higher priority than ATIs, they contain more information that has to be recorded and verified by the FIST HQs. Therefore, it was not surprising that the median time (13.5 seconds) for fire requests was 55 percent higher than the median service time (8.5 seconds) for ATIs.

From Table 6, which considers both fire requests and ATIs, it is obvious by examining the marginals of this table that an increase in intensity or degradation resulted in an increase in the FIST's service time. There was a 37 percent increase in median service time as degradation increased from 0 to 30 percent and a 37 percent increase in median service time as intensity increased from low to high. However, the effect that intensity had on the FIST HQs service time is not as predominant or does not exist when considering ATIs and fire requests separately. The median service time for ATIs increased 12 percent from low to high intensity as observed in the right marginal of Table 8. Contrary to this trend, the FIST HQ's ability to service fire requests remained essentially the same in either low or high intensity as shown in Table 7. One possible explanation is that as intensity increased, more effort was made to service the fire request messages that have a higher priority than ATIs. Consequently, ATIs were not serviced as quickly.

The effect that degradation had on service time is consistent with the above trend for both ATIs and fire requests. As observed in examining the bottom marginals of Tables 7 and 8, an increase in degradation from 0 to 30 percent resulted in the FIST HQ's median service

**TABLE 5. Analysis of Variance
(Effect on Service Time)**

SOURCE	DEGREES OF FREEDOM	SUM OF SQUARES	MEAN SQUARE	F RATIO
Replication	1	5.62	5.62	8.64**
Message Type	1	6.01	6.01	9.25**
Block within Rep	4	16.45	4.11	
Message Type X Block within Rep	4	7.36	1.84	
Team	1	150.98	150.98	232.3**
Team X Block within Rep	4	17.1	4.28	
Intensity	2	14.77	7.38	11.2**
Intensity X Message Type	2	9.48	4.74	7.18**
Intensity X Team	2	7.25	3.63	5.5**
Degradation	2	52.68	26.34	39.91**
Degradation X Message Type	2	5.52	2.76	4.18**
Degradation X Team	2	2.60	1.30	1.96
Intensity X Degradation	4	31.68	7.92	12.01**
Intensity X Degradation X Team	4	11.99	3.00	4.54**
Pooled Error	790	520.9	.66	
Total	825	860.39		

time increasing 29 percent and 13 percent for fire request and ATIs, respectively.

**TABLE 6. Intensity by Degradation
Median Service Time
Fire Requests and ATIs**

Intensity	Communication Degradation (%)			
	00	15	30	
LOW	9.0	9.5	9.0	9.5
MEDIUM	9.5	11.0	11.0	10.5
HIGH	10.5	12.5	18.0	13.00
	9.5	11.0	13.0	

The ANOVA table showed a significant interaction intensity degradation effect on service time. As observed in Table 6, this trend was slight in low or medium intensity as degradation increased from 0 to 30 percent. However, in high intensity, the increase from 0 to 30 percent degradation resulted in a 71 percent increase in service time which was substantially higher than the increases observed in low or median intensity as degradation increased. This interaction effect was prevalent for both ATIs and fire requests.

For ATIs, the median service time increased only slightly as degradation increased from 0 to 30 percent for low or medium intensity as shown in Table 8. Similarly, for fire request messages, the increase in degradation from 0 to 30 percent was only 4 percent in medium intensity. This trend was more noticeable in low intensity where the median FIST HQ's service time for fire requests increased almost 28 percent as degradation increased from 0 to 30 percent. However, in high intensity, the increase from 0 to 30 percent degradation resulted in a substantial increase in service time for both ATIs and fire request messages when compared to any increase observed in low or medium intensity. The median service time for fire requests increased 46 percent from 0 to 30 percent degradation and for ATIs increased 179 percent. This was due to the fact that the largest median service time observed

**TABLE 7. Intensity by Degradation
Median Service Time
for Fire Requests**

Intensity	Communication Degradation (%)			
	00	15	30	
LOW	12.5	13.5	16.0	14.5
MEDIUM	11.5	13.5	12.0	12.0
HIGH	12.0	13.0	17.5	14.0
	12.0	13.0	15.5	

for ATIs and fire requests occurred under 30 percent degradation and high intensity. In addition, it was only under this condition that the median service time (19.5 seconds) for ATIs was higher than for the median service time (17.5 seconds) for fire requests. This seems to substantiate the hypothesis that under increased workload, the FIST HQs spends more time trying to service fire request messages while ATIs are left in the DMD queue.

Although replication (learning) was significant, only a slight decrease (8 percent) in service time was observed between replicate 1 and replicate 2.

The final step in this analysis was to categorize the removed data by various experimental conditions. The following trends were worth noting. Of the 36 service times removed from the data base, over one third were observed under 30 percent degradation and high intensity. In addition, 75 percent were observed from 30 percent degradation with over 92 percent coming from two hour cells that were run under 15 or 30 percent degradation. These observations substantiate that increased degradation and the combined effect of 30 percent degradation and high intensity caused delays for the FIST HQs in servicing messages.

**TABLE 8. Intensity by Degradation
Median Service Time
for ATIs**

Intensity	Communication Degradation (%)			
	00	15	30	
LOW	8.5	8.5	8.5	8.5
MEDIUM	7.5	9.5	9.0	9.0
HIGH	7.0	9.5	19.5	9.5
	8.0	9.0	9.0	

IV. CONCLUSIONS

Software, intensity, communication degradation, software-intensity interaction and intensity-degradation interaction all have a significant effect on message traffic through the FIST HQs. A change from 0 to 15 percent communication degradation resulted in an average increase of 33 percent in the number of messages generated. A change from 0 to 30 percent communication degradation resulted in an average increase of 73 percent in the number of messages generated. Medium intensity generated 75 percent more messages than low intensity and high intensity generated 144 percent more messages than low intensity, on the average. Software was added as a factor in the experiment to control for the variance induced by the change in software. Knowing that the change was significant and the second set of software was more correct tactically, only the second half of the test was analyzed for the other measures.

The number of transmissions of a given message before an acknowledgement is received is important because the FIST DMD allows only four tries and then voice contact must be made to synchronize authenticator codes. Voice transmissions on digital nets cause contention. In 15 percent degradation .2 percent of the messages required more than four transmissions and in 30 percent degradation 2.4 percent of the messages required more than four transmissions. Although these percentages are small, because of the large number of messages on any net the actual number of voice transmissions required may be tactically significant.

The median service time for messages was influenced significantly by team, message type, replication, intensity, degradation, and many of the interaction terms. It is not surprising that when measuring a human response time that the humans, or FIST HQs, are the most significant factor. Replication being significant in this instance can be translated to a slight learning effect since the first replicate occurred on the first three days of testing and the second replicate occurred on the last three days. An increase of 32 percent in median service time for fire requests and ATIs combined was observed from 0 to 30 percent degradation and an increase of 34 percent was observed as intensity increased from low to high. The combined effect of intensity degradation is most noticable in high intensity. That is, communication degradation has little effect within low intensity or medium intensity, but has a very large effect in high intensity. Because intensity is defined by weighing the initiating message types (fire requests and ATIs) when we breakout service time by message type, we no longer observe the effect of intensity. What we do notice, however, is that although fire requests take longer to process in general than ATIs, as communication degradation increases within high intensity the rate at which service time increases for ATIs is considerably higher than the rate of increase for fire requests until finally at 30 percent degradation ATIs take longer to process than fire requests. Service time in high intensity increases 179 percent for ATIs and 46 percent for fire requests. What this would indicate is a queueing problem at the FIST HQs. Fire requests are higher priority than ATIs and are selected out of the queue before ATIs for processing. Therefore, although it may not take as long to process ATIs they are remaining in the queue longer until finally their service time exceeds that of fire requests because service time is both the time spent in the FIST DMD queue and the human processing time.

At the time this paper was presented, the complete analysis of the data produced was not completed and is, therefore, not presented in these proceedings. Complete analyses will be published in a BRL report upon completion and can be requested from the authors.

APPENDIX A

General Analysis of Variance Procedure

Analysis of variance (ANOVA) is a common procedure which tests the hypothesis that there is no statistical difference between the mean value of data drawn from two or more populations. One can think of a population as data collected under the same experimental conditions. This procedure utilizes the F-statistic which is a ratio of the estimated variance (mean square) of the factor or interaction one is testing divided by its associated error. The number of factors evaluated and their associated error is dependent upon the model one chooses, and if the different factors in the model are fixed or random. A factor is considered random if it contains categories or levels which are considered samples from a larger group. A fixed factor is one in which its categories or levels exhaust the cases in which there is interest. Also, the categories are not merely samples.

Corresponding to the F-statistic is a significance level $(1-\alpha)$ where α is the probability of rejecting a true hypothesis. For this analysis, α will be equal to .05. If the calculated F-statistic is larger than the tabulated F-value, then the hypothesis that the factor has no effect in a given measure of performance (MOP), is rejected. However, the test of significance using the F distribution is valid if the observations (MOPS) are from normally distributed populations with equal variances. Investigation has shown that results of the analysis are robust to the departure from the assumption of normal distribution but the homogeneity of variance assumption should be checked.

The model on which our analysis is based contained all possible treatment combinations as specified by the design except interaction terms that contained replication. If only one observation per experimental condition was available, the interaction terms containing replication were assumed not to be significant and were included in the estimate of error. If more than one observation per cell was available, then Bartlett's test was performed on the these cells sample variances. If these mean squares or variances were found to be different, then an appropriate transformation was performed on the MOP being evaluated so that these estimates can be used as the error term in the model.

Based on the above described model and the fact that replication was the only term considered random, the expected mean squares were determined as shown in Table A-1.

Examining the components of the error mean square, the F-ratio can be determined for each treatment combination. For example, the expected mean square for replication contains a source of variation for replication and pooled error. Therefore, the proper denominator for the F-ratio is the mean square for error. Due to the model specifications, the pooled error term happens to be the proper denominator to use for every term in the model. The estimated mean square for each treatment combination is obtained by calculating each effects sum of squares and dividing by its associated degrees of freedom.

The degrees of freedom used to calculate mean squares for each treatment combination are given in Table A-2. By comparing the calculated F-statistic to the tabulated F-value one can determine if each treatment combination had an effect on fire support coordination based on the MOP being evaluated.

TABLE A-1. Analysis of Variance
(Expected Mean Squares)

ANALYSIS OF VARIANCE (ANOVA)	
SOURCE	EXPECTED MEAN SQUARE
Replication	$108 \cdot \sigma_r^2 + \sigma_e^2$
Software	$108 \cdot \phi_s + \sigma_e^2$
Block within Rep	$360 \cdot \phi_b + \sigma_e^2$
Software X Block within Rep	$18 \cdot \phi_{sb} + \sigma_e^2$
Team within Software	$54 \cdot \phi_t + \sigma_e^2$
Team within Soft X Block within Rep	$9 \cdot \phi_{tb} + \sigma_e^2$
Intensity	$72 \cdot \phi_i + \sigma_e^2$
Software X Intensity	$36 \cdot \phi_{si} + \sigma_e^2$
Intensity X Team w Software	$18 \cdot \phi_{it} + \sigma_e^2$
Degradation	$72 \cdot \phi_d + \sigma_e^2$
Software X Degradation	$36 \cdot \phi_{sd} + \sigma_e^2$
Team within Software X Degradation	$18 \cdot \phi_{td} + \sigma_e^2$
Intensity X Degradation	$24 \cdot \phi_{id} + \sigma_e^2$
Soft X Intensity X Degradation	$27 \cdot \phi_{sid} + \sigma_e^2$
Team within Software X Intensity X Degradation	$60 \cdot \phi_{tid} + \sigma_e^2$
Error	σ^2

**TABLE A-2. Analysis of Variance
(Degrees of Freedom)**

ANALYSIS OF VARIANCE (ANOVA)	
SOURCE	DEGREES OF FREEDOM
Replication	1
Software	1
Block within Rep	4
Software X Block within Rep	4
Team within Software	2
Team within Soft X Block within Rep	8
Intensity	2
Software X Intensity	2
Intensity X Team w Software	4
Degradation	2
Software X Degradation	2
Team within Software X Degradation	4
Intensity X Degradation	4
Soft X Intensity X Degradation	4
Team within Software X Intensity X Degradation	8
Pooled Error	19
Total	71

The final step in the ANOVA procedure is that the interaction terms found not to be significant can also be pooled with the error component of the model and the analysis redone. This procedure reduces the model and increases the degrees of freedom for error and subsequently increases the confidence in conclusions reached if both analyses agree.

APPENDIX B

(Contingency Table Analysis)

Contingency table analysis is a method used to make direct inferences about whether two or more population distributions are identical to some theoretical form. Ordinarily, the reason for comparing such distributions is to find evidence for independence of attribute or experimental conditions. In short, we are going to employ a test for independence for each experimental unit in our design matrix. The general procedure is to statistically compare the sample or observed frequency for each experimental unit to the theoretical expected frequency.

The statistic used to test if the observed frequency for each treatment combination is equal to the expected frequency is the chi-square statistic. This statistic is defined as

$$\sum_{i=1}^{i=N} \frac{(f_i - e_i)^2}{e_i}$$

where N is the number of experimental units and f_i and e_i are the observed and expected cell frequencies. The calculated statistic is then compared to a tabulated value which is based on an alpha level equal to .05 and the number of degrees of freedom associated with the analysis. The number of degrees of freedom is equal to the number of experimental unit minus one, minus the number of parameters estimated from the sample data which are needed to determine the expected frequency. If the calculated chi-square statistic is larger than the tabulated value, the hypothesis that the experimental treatments are not associated with the MOP being analyzed is rejected. One restriction is that the sample size must be sufficiently large so that none of the theoretical frequencies are less than 1 and not more than 20 percent are less than 5.

For MOP4, which is the frequency count of the number of times a message is sent before it is acknowledged, the theoretical distribution can be determined for each treatment combination without any sample results. At zero percent degradation, the probability of having a try number greater than zero, which can be interpreted as the probability of a message not getting through and/or an acknowledgement not being returned on the first try, is zero. At fifteen percent degradation the probability of a message getting through and an acknowledgement returned on the first try is recorded for each two-hour block run with 15% degradation to have a try number of zero. Similarly, with 30% degradation, one would expect 49 percent of the total messages recorded per two hour block to have been tried only once. The theoretical probability by try is given in Tables B-1 and B-2.

If needed three separate contingency analyses will be performed for each level of communication degradation. However, at zero percent degradation all of the messages should be acknowledged after the first try. The expected number of messages by try number for the 24 cells run at each level of communication degradation is presented below. It is worth noting that since no parameter estimation is needed to determine these theoretical distributions the degrees of freedom for each analysis is equal to the number of cells minus one. These theoretical frequencies were compared to the observed frequencies using the

TABLE B-1. Probability of a Message Being
Acknowledged by Try
(15% Degradation)

Degradation	Try			
	1	2	3	greater 3
15 %	.723	.201	.056	.021

TABLE B-2. Probability of a Message Being
Acknowledged by Try
(30% Degradation)

Degradation	Try			
	1	2	3	greater 3
30 %	.490	.250	.128	.132

above described procedure. Then, using contingency table analysis outlined above, one can determine if the other experimental factors had an effect on the number of tries it takes before a message is acknowledged.

The first step of this analysis was to verify that the uniform number generator did produce fifteen and thirty percent total message lost for each set of twelve cells run under the modified software. Using the chi-square statistic defined above, one can test if in fact the observed and expected number of messages never degraded under 15 and 30 percent degradation are statistically the same.

For fifteen percent degradation, the chi-square statistic was calculated as 2.257 with 11 degrees of freedom and found not significant at alpha equal to .05. Similarly, at 30 percent degradation, the statistic was calculated as 1.175 with 11 degrees of freedom and again found not to be significant. In fact, over each set of twelve cells, it was calculated that .8525 and .7054 of the messages were never degraded for 15 and 30 percent degradation, respectively.

Having verified that ETHER was producing the desired degradation levels in our communication network, the next step is to determine if intensity and team variability had an effect on the distribution of message tries for acknowledged messages at each degradation level.

At 0 percent degradation, one would expect all of the messages to be acknowledged on the first try. As seen from Table B-3 below, almost all (98.6%) of the messages had successfully been sent and acknowledged. It is obvious that intensity and team variability had no effect on a message reaching its destination at zero percent degradation. The 1.4 percent

**TABLE B-3. Observed Number of Messages Acknowledged
by Try
(00% Degradation)**

Rep	Software	Team	Intensity	Try		
				1	2	3
1	2	3	L	134	9	1
			M	192	3	0
			H	265	0	0
		4	L	121	1	0
			M	201	3	0
			H	277	3	0
2	2	3	L	112	2	0
			M	192	4	0
			H	269	2	0
		4	L	116	2	0
			M	198	2	0
			H	271	2	0

of the messages that did not get through on the first try can be attributed to bit box collisions which is a hardware phenomena. This phenomena occurs when two messages enter the bit box on opposite ends simultaneously, collide and then are lost.

A contingency table analysis was performed on the 12 two-hour cells run at 15 percent communication degradation. The observed number of messages acknowledged for try one, two, three and tries greater than three were compared to the expected number. The calculated chi-square statistic was 44.2 with 47 degrees of freedom. This statistic was not statistically significant and one can only conclude that the observed and theoretical distributions are the same.

For 30 percent degradation the contingency table analysis again revealed that intensity, team variability and replication did not influence the number of tries it took for a message to be acknowledged. The chi-square statistic was 30.29 with 59 degrees of freedom.

In conclusion, based on our experiment, we have demonstrated that the number of tries it took before a message was acknowledged is a function of the percent degradation exhibited in the communications network and it is not statistically influenced by intensity, team variability or replication. The theoretical and actual frequency distributions by try number are given in Tables B-4 through 7 for 15 and 30 percent communication degradation, respectively.

TABLE B-4. Observed Number of Messages
Acknowledged by Try
(15% Degradation)

Rep	Software	Team	Intensity	Try				
				1	2	3	greater 3	
1	2	3	L	96	29	5	3	
			M	153	57	17	2	
			H	194	47	18	6	
		4	L	76	35	6	3	
				M	166	41	11	3
				H	197	69	12	6
2	2	3	L	92	22	2	3	
			M	143	48	8	4	
			H	197	71	14	7	
		4	L	88	37	12	5	
				M	150	45	10	8
				H	206	61	13	4

**TABLE B-5. Observed Number of Messages
Acknowledged by Try
(15% Degradation)**

Rep	Software	Team	Intensity	Try			
				1	2	3	greater
1	2	3	L	96	26.7	7.5	2.7
			M	165.5	46.0	12.8	4.7
			H	191.5	53.3	14.8	5.4
		4	L	86.7	24.1	6.7	2.5
			M	159.7	44.4	12.4	4.5
			H	205.2	57.1	15.9	5.8
2	2	3	L	86	24	6.7	2.4
			M	146.6	40.1	11.4	4.2
			H	208.8	58.1	16.2	5.9
		4	L	102.6	28.5	7.9	2.9
			M	153.9	42.8	11.9	4.4
			H	205.2	57.1	15.9	5.8

TABLE B-6. Observed Number of Messages
Acknowledged by Try
(30% Degradation)

Rep	Software	Team	Intensity	Try				
				1	2	3	4	greater 4
1	2	3	L	68	34	26	8	12
			M	94	71	29	13	12
			H	154	63	41	23	19
		4	L	67	34	14	8	8
			M	106	62	30	14	12
			H	150	73	33	13	18
2	2	3	L	64	23	18	11	10
			M	115	52	32	10	16
			H	150	74	43	18	17
		4	L	65	34	15	7	9
			M	111	56	24	14	14
			H	161	101	32	27	22

TABLE B-7. Expected Number of Messages
Acknowledged by Try
(30 % DEGRADATION)

Rep	Software	Team	Intensity	Try				
				1	2	3	4	greater 4
1	2	3	L	72.6	37	18.9	9.6	10
			M	107.4	54.7	27.9	14.2	14.8
			H	147	75	38.3	19.5	20.3
		4	L	64.2	32.7	16.7	8.5	8.8
			M	109.8	56	28.6	14.6	15.1
			H	140.6	71.7	36.6	18.7	19.4
2	2	3	L	61.7	31.5	16.1	8.2	8.5
			M	110.3	56.2	28.7	14.6	15.2
			H	148	75.5	38.5	19.6	20.4
		4	L	63.7	32.5	16.6	8.5	8.8
			M	107.3	54.7	27.9	14.2	14.8
			H	168.1	85.7	43.7	22.3	23.1

A TECHNIQUE TO APPROXIMATE COMPLEX COMPUTER MODELS

An Approximation of the Teisberg Model

Joseph M. Tessmer

This paper was prepared while the author was employed by the Office of the Strategic Petroleum Reserve within the Department of Energy.

The author is currently assigned to:

Headquarters, US Air Force
ACS/Studies and Analyses
Fighter Division
Washington, DC 20330

Telephone: 202-697-5677
A/V 227-5677

Abstract

This paper presents a technique which was used to produce an approximation of a complex computer model, the Teisberg Model. The technique employs a complete 2^4 factorial design and uses the statistically significant effects as coefficients of the estimating equation.

Disclaimer

The assumptions, procedures, analysis, conclusions, and recommendations contained in this paper are solely those of the author and do not represent any official policy of the Department of Energy, the Department of Defense, or US Government.

An Approximation of the Teisberg Model

Background

An approximation was constructed of the Teisberg model which estimates the economic benefit of constructing and maintaining a Strategic Petroleum Reserve. Four input factors, which replicated significant independent economic assumptions were identified as candidates for inclusion within the simplified model. The four variables were:

1. p = Annual probability of a major oil disruption
2. e = The short run price elasticity of demand for oil
3. b = The BAU price of crude oil
4. d = The discount rate

Using a one variable at a time approach three of these variables were set at the center, of their range of interest, and the Teisberg Model estimated the net economic benefit (Y) for a low, medium, and high value of the remaining variable.

This was done for the four candidate variables. An estimate of the rate of percent of change of the economic benefit Y to the percent change of the input factor X was calculated i.e., $\frac{dY/Y}{dX/X}$.

The results of this effort were:

Input factor	$\frac{dY/Y}{dX/X}$
Probability of a major disruption	0.543
Short run price elasticity	-2.196
BAU price of crude oil	0.330
Discount rate	-0.864

It was determined that only the short run price elasticity for demand need be considered when estimating the results of the Teisburg Model.

A linear regression was then performed on the three observations of the Teisberg Model with the low, medium, and high values for the elasticities and the three remaining variables set at the center of their range of interest.

The resulting equation was $Y = 275.85 e + 83.67$ where e is the elasticity of demand, $-0.3 < e < -0.1$ and Y is the estimate of net economic benefit. The R^2 value was 0.86 which seems to indicate a good approximation. However, only three observations were used and two are required to determine a straight line, leaving only one degree of freedom, and thus a high R^2 .

The Alternate Estimate

At the request of the principal investigator the sound principles of experimental design were applied to the same problem with the hope that an improvement might be made in the estimating equation. The remainder of this paper and the appendixes are the result of that request.

The estimate of the net economic benefit using the techniques of experimental design is:

$$y = 119.57 - 1137.87d + 398.00 e + 2148.00 p$$

$$- 4216.00de - 7488.00 dp + 5246.00 ep \quad \text{where}$$

d = discount rate $0.025 < d < 0.1$

e = elasticity of demand $-0.3 < e < -0.1$

p = annual probability of a major disruption $0 < p < 0.1$.

Details of the theory and construction of this estimate appear in the appendixes. The relative merits of the two estimates may be established by examining the estimates of both equations using the observations used in this study.

<u>Observation</u>	<u>Teisberg Value</u>	<u>Original Estimate</u>	<u>Estimate Residual</u>	<u>Alternate Estimate</u>	<u>Estimate Residual</u>
1	3.48	0.92	2.56	12.86	-9.38
2	14.69	0.92	13.77	3.34	11.35
3	2.56	0.92	1.64	12.86	-10.30
4	15.72	0.92	14.80	3.34	12.38
5	18.71	56.09	-37.38	8.14	10.57
6	50.06	56.09	-6.00	61.86	-11.80
7	17.28	56.09	-38.83	8.14	9.14
8	49.97	56.09	-6.12	61.86	-11.89
9	7.95	0.92	7.03	0.40	7.55
10	27.01	0.92	28.09	47.04	-20.03
11	27.01	0.92	11.60	0.40	12.12
12	43.39	0.92	42.47	47.04	-3.65
13	67.62	56.09	11.53	100.60	-32.98
14	169.16	56.09	113.07	210.48	-41.32
15	113.90	56.09	57.81	100.60	13.30
16	275.48	56.09	219.37	210.48	65.00
Sum of squared residuals $\sum (Y - \hat{Y})^2$			70,564.85		8,767.37
mean square error $\sum (Y - \hat{Y})^2 / 16$ (unadjusted for degrees of freedom)			4410.30		547.96

Table 1

Caveat

This estimate or approximation of the Teisberg Model was based on assumptions for several input factors which were not varied during this exercise. Changes in the values for these input factors may alter the quality of this estimate.

Next Steps

There are some promising techniques that may lead to additional improvements in an estimate of the Teisberg Model. The first is the application of response surface analysis to estimate the coefficient of higher ordered terms. The second involves various transformations, of the data, as the first step of the analysis. Thirdly, additional input factors might be included in the analysis. These techniques used independently, or in conjunction with each other, should improve the quality of the estimate.

Appendix A METHODOLOGY

A1 Factorial Design Methodology

An experiment was performed to measure the effect of four sets of input factors on the average net economic benefit associated with four SPR alternatives, as represented by the Teisberg model. Two levels, for each set of input factors, were chosen and all 16 possible combinations of these input factors, were used as model input to the Teisberg model. This procedure, a 2^4 factorial design was chosen since it is economical, easy to use and provides a great deal of valuable information. Specifically a two (2) level factorial design has the following advantages:

1. If sets of input factors are varied one set at a time, with the remaining factors held constant, it is necessary to assume that the effect would be the same at other settings of the other sets of input factors. Factorial designs avoid this assumption.
2. If the effects of input factors act additively, a factorial design estimates those effects with more precision. If the effects of the input factors do not act additively, factorial designs can detect and estimate the interactions which measures the non-additivity.
3. Factorial designs require relatively few runs per set of input factors studied and can indicate major trends and determine promising direction for further investigation. To obtain the same precision of the estimate of the effects measured, in this effort, forty runs would have had to be run, using the traditional, one factor at a time approach, rather than the sixteen used in the experiment.
4. If a more thorough local exploration is needed, it can be suitably augmented to form composite designs.
5. These designs and their corresponding fractional designs may be used as building blocks so that the degree of complexity of the finally constructed design can match the sophistication of the problem.

To perform a 2^4 factorial design the two extreme levels (or versions), as defined by the principle investigator, were selected for the four (4) sets of input factors and all sixteen (16) possible combinations were run, which created sixteen observations. The four sets of input factors and their levels (or versions) are listed in Table A-1 on the following page.

Input Factor

Levels

1, Probability of a major oil disruption	1a, 0.0, no chance of a major oil disruption during any year of the study. 1b, 0.1 A ten percent chance in any given year of a major oil disruption
2, The short run price elasticity	2a, - 0.3 a low short run elasticity of demand for oil 2b, - 0.1 a high short run elasticity of demand for oil
3, The business as usual price for crude oil	3a, \$52.00 per barrel, a low price 3b, \$90.00 per barrel, a high price
4, The discount rate	4a, 10.0% the conventional government discount rate 4b, 2.5% a low discount rate

TABLE A-1

The selection of the above levels were determined by the parent study and do not represent the policy of the Department of Energy. These levels were used solely to evaluate the reaction of the Teisberg Model to changes in the input factors.

These input factors combine to produce the following design matrix:

Design Matrix

<u>OBS.</u> <u>NUMBER</u>	<u>PROB.</u> <u>DISRUPT</u>	<u>ELAS.</u>	<u>PRICE</u> <u>CRUDE</u>	<u>DISCOUNT</u> <u>RATE</u>	<u>TEISBERG</u> <u>NET BEN.</u>
1	1a	2a	3a	4a	3.48
2	1a	2a	3a	4b	14.69
3	1a	2a	3b	4a	2.56
4	1a	2a	3b	4b	15.72
5	1a	2b	3a	4a	18.71
6	1a	2b	3a	4b	50.00
7	1a	2b	3b	4a	17.28
8	1a	2b	3b	4b	49.97
9	1b	2a	3a	4a	7.95
10	1b	2a	3a	4b	27.01
11	1b	2a	3b	4a	12.52
12	1b	2a	3b	4b	43.39
13	1b	2b	3a	4a	67.62
14	1b	2b	3a	4b	169.16
15	1b	2b	3b	4a	113.90
16	1b	2b	3b	4b	275.48

Table A-2

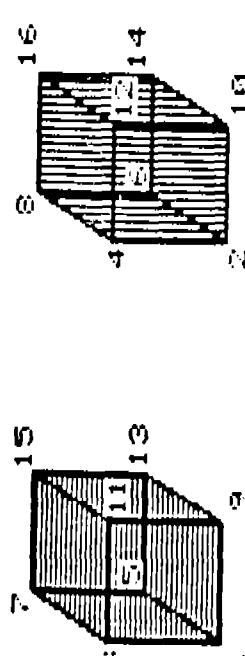
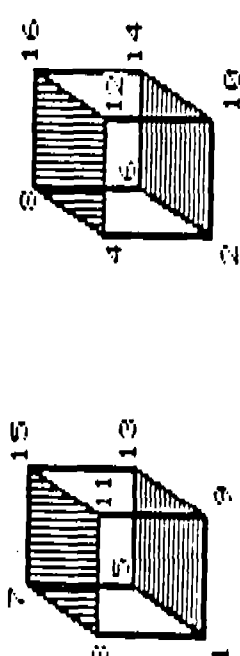
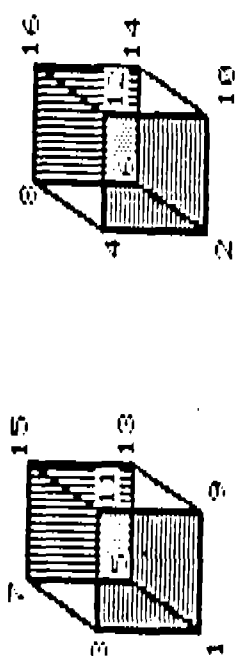
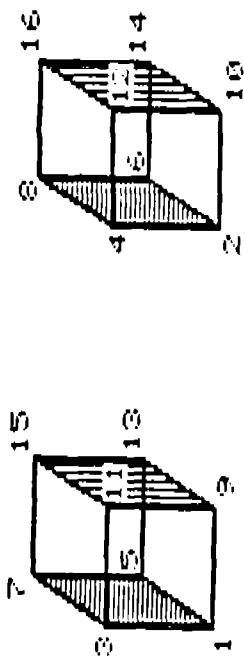
The interpretation of the observations in Table A-2 is easily illustrated by observation number 6 which assumes that the annual probability of a major oil disruption is 0.0 i.e. there will not be a major disruption during this study. There is a high elasticity of demand for crude oil of -0.1 with a business as usual price for crude oil of \$52.00 per barrel. Finally a low discount rate of 2.5% is assumed.

The sixteen observations of the design matrix, may be visualized geometrically as two cubes. One possible visualization appears in figure A-1 on the following page. The observation number is at each vertex.

THE TEISBERG MODEL

Figure A-1

GEOMETRICAL REPRESENTATION OF THE 2^4 FACTORIAL DESIGN MAIN EFFECTS



PROBABILITY OF A
MAJOR OIL DISRUPTION

DEMAND ELASTICITIES

B.R.U. CRUDE PRICE

DISCOUNT RATE

A2 Calculation of Main Effects

The "main effect" of a set of input factors is the change in the response i.e., the net economic benefit, y , as we move from the "a" case to the "b" case version of that set of input factors. To examine the effect of each of the selected input factors a table of four column vectors was constructed (see table A-3). Each column contrasts eight pairs of estimates of the net economic benefit. Aside from experimental error, the difference between the upper number of a pair and the lower number of the same pair is due to the change of the input factor that heads the column. For each column the average of these eight differences is the main effect due to the associated input factor that heads the column. Note that the only difference between the four columns is the order in which the observations appear.

Geometrically speaking, using Figure A-1 the main effects are calculated from the corresponding vertices of the two cubes as described below.

Input factor

Probability of a major oil
disruption

Left side of both cubes vs.
the right side of both
cubes

Demand elasticities

The front of both cubes vs.
the backs of both cubes

Business as usual crude price

The bottom of both cubes
vs. the tops of both cubes.

Discount rate

The left cube vs. the right
cube.

Main Effects
Table of Contracts

<u>Prob. of major oil disruption</u>		<u>Demand Elasticities</u>		<u>BAU Crude Price</u>		<u>Discount Rate</u>	
<u>Obs. Number</u>	<u>Net Econ. Benefit</u>	<u>Obs. Number</u>	<u>Net Econ. Benefit</u>	<u>Obs. Number</u>	<u>Net Econ. Benefit</u>	<u>Obs. Number</u>	<u>Net Econ. Benefit</u>
1	3.48	1	3.48	1	3.48	1	3.48
9	7.97	5	18.71	3	2.56	2	14.69
2	14.69	2	14.69	2	14.69	3	2.56
10	27.01	6	50.06	4	15.72	4	15.72
3	2.56	3	2.56	5	18.71	5	18.71
11	12.52	7	17.28	7	17.28	6	50.06
4	15.72	4	15.72	6	50.06	7	17.28
12	43.39	8	49.97	8	49.97	8	49.97
5	18.71	9	7.97	9	7.97	9	7.97
13	67.62	13	67.62	11	12.52	10	27.01
6	50.06	10	27.01	10	27.01	11	12.52
14	169.16	14	169.16	12	43.39	12	43.39
7	17.28	11	12.52	13	67.62	13	67.62
15	113.90	15	113.90	15	113.90	14	169.16
8	49.97	12	43.39	14	169.16	15	113.90
16	275.48	16	275.48	16	275.48	16	275.48

TABLE A-3

A3 2nd-Order Interaction Effects

Suppose that one is interested in examining the effects of two sets of input factors; for example, the probability of a major interruption and the discount rate. Then the sixteen runs of the factorial design can be grouped into four sets of four runs each. Each run in the group would have the same value for the input factors studied, although other input factors would vary within each group. Assume that if there is no chance for a major oil disruption and the discount rate is 10%, that the average value for the output variable being studied is 100. This will be the base point. Also assume that the main effects for the probability of a major interruption and the discount rate are 25 and 10 respectively. This means that, on the average, changing from no chance of a major interruption to an annual probability of an interruption of 0.10 will increase the output variable under study by 25. Likewise a change in the discount rate from 10% to 2.5%, will on the average, increase this same output variable by 10. If the input factors act additively, then the average value of the output variable with 0.10 chance of an interruption and a 2.5% discount rate would be $100 + 25 + 10 = 135$.

This artificial case is represented by the upper diagram in figure A-2. Note that the quantity

$$(b + c - a - d)/2 = (110 + 125 - 100 - 135)/2 = 0$$

i.e., there is no interaction.

Suppose that the input factors do not act additively, and the base point of 100 and main effects are the same. Then the resulting measurements could be described by the lower diagram in figure A-2. The input factors are now said to interact. By convention a measure of this interaction is

$$(b + c - a - d)/2 = (145 + 160 - 100 - 135)/2 = 35$$

This is a second order interaction and is called the probability of a major oil interruption X discount rate interaction.

Like a main effect, a 2nd order interaction is the difference between two averages, eight of the sixteen results being included in one average and eight in the other. Analogous explanations are easily constructed for all other 2nd order interaction effects.

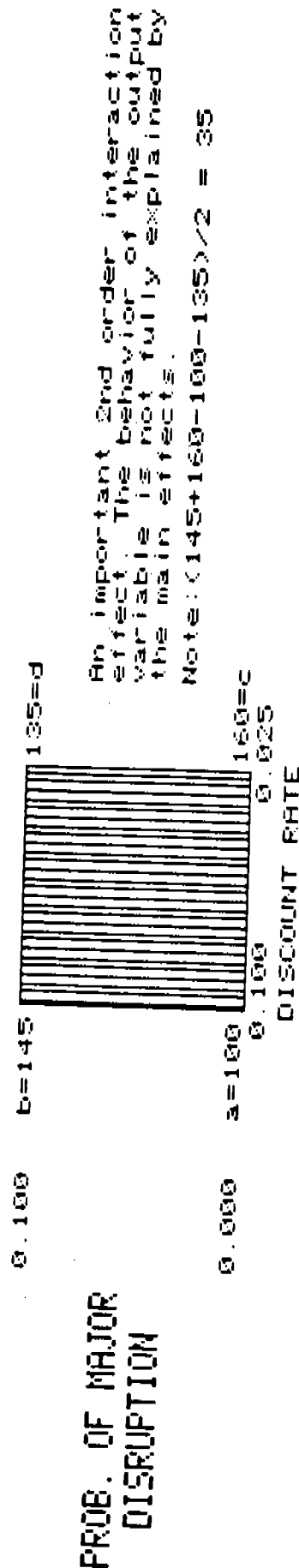
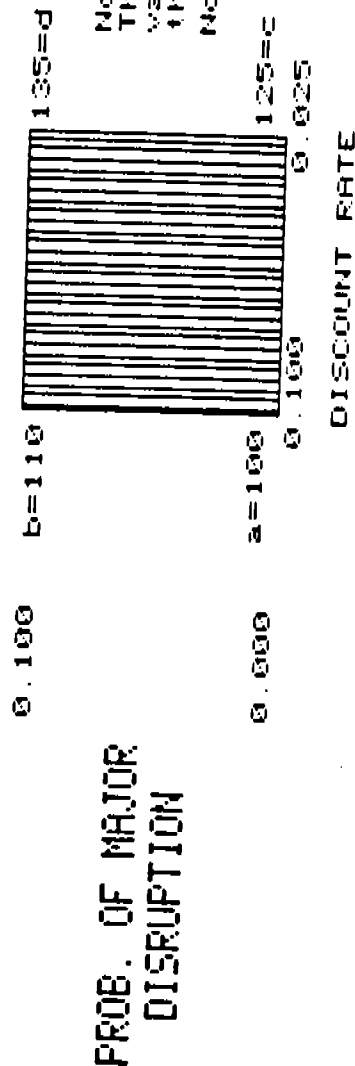
A4 Higher-Order Interaction Effects and the Standard Error.

Similar procedures to those above can be given for deriving the third and fourth-order interactions. Due to the similarity of response functions it is reasonable to assume that higher-ordered interactions are negligible and measure differences arising principally from experimental error. Thus the mean, of the sum of squares, of these interactions give an estimated value for the variance of an effect, having five degrees of freedom. The square root of this value is an estimate of the standard error.

THE TEISBERG MODEL

figure A-2

Interpretation of 2-Way Diagrams



The level of statistical significance chosen for this study was $p=0.10$. In order to select the statistically significant main effects and second order interactions multiply the standard error by $t_{1-p/2}=2.015$. Any main effect or interaction with absolute value greater than this product is considered statistically significant.

A5 The Plot of Effects

If the output from the model had simply occurred by chance, the observations would be normally distributed about some fixed mean, and the changes in the input factors would not have a real effect on the estimate of the net economic benefit. The fifteen effects, main effects plus all interactions, could then be plotted on normal probability paper as straight line. One may conclude that the effects that are not roughly on this straight line, are due to changes in the input factors and have a significant effect on the output variable being studied.

A6 The Binary Estimates

Define $X_i = \begin{cases} -1 & \text{if } ia \text{ is the value of the } i \text{ th input factor} \\ & \text{(see table A-1).} \\ 1 & \text{if } ib \text{ is the value of the } i \text{ th input factor} \\ & \text{(see table A-1).} \end{cases}$

Let a_i be the main effect of the i th input factor

Let a_{ij} be the 2nd order interaction of the i th and j th input factors.

Let I index the set of significant main effects at a fixed level of significance p .

Let IJ index the set of significant 2nd order interactions at the same fixed level of significance. The binary 2nd order estimates of the process is

$$Y = Y + \sum_{i \in I} (a_i/2) X_i + \sum_{ij \in IJ} (a_{ij}/2) X_i X_j$$

A7 The Residual Plot

If the number of significant effects is small compared to the total number of residuals then one can interpret the plot of residuals on normal probability paper. If the residual points lie more or less on a straight line then one may conclude that the unexplained variation is due to random noise and that the identified significant effects explain the process. If this does not happen then the proposed binary estimate does not fully capture the underlying process and more work needs to be done.

A8 The Continuous Estimate

If an input factor, is in fact a continuous variable, with an interval or ratio scale, then the binary estimate may be transformed to a continuous estimate. Let z_1 be the continuous input factor such that:

$$z_1 = \begin{cases} ia & \text{in the a case} \\ ib & \text{in the b case} \end{cases}$$

Note that $X_1 = (2z_1 - ia - ib)/(ib - ia)$

has the following property:

a, if $z = ia$ then $X_1 = -1$

b, if $z = ib$ then $X_1 = 1$

To construct the continuous estimate replace X_1 in the binary estimate with $(2z_1 - ia - ib)/(ib - ia)$.

Appendix B APPLICATION

B1 Analysis of the Net Economic Benefit

The main effects of three of the input factors, the discount rate, the demand elasticities and the probability of a major disruption are statistically significant at the $p < .10$ level. In addition there are perceptible 2nd order interactions between each pair of the input factors which had statistically significant main effects. Therefore each pair of these input factors must be evaluated jointly. The two way diagram of figure B-1 depicts the nature of these interactions.

Assuming a conventional discount rate of 10% the Teisberg Model estimates that an increase of the BAU price of crude oil from \$52.00 per barrel to \$90.00 per barrel will increase the net economic benefit from \$6.63 billion to \$54.38 billion. If a discount rate of 2.5% is assumed, the identical change in the price of crude oil will increase the net economic benefit from \$25.20 billion to \$136.17 billion.

Given the assumption that there is virtually no chance of a major disruption the Teisberg Model estimates that a change of the discount rate from 10.0% to 2.5% will increase the net economic benefit from \$10.51 billion to \$32.61 billion. If the annual probability of major disruption is 0.10 then the identical change in the discount rate increase the probability of a major disruption from \$50.50 billion to \$128.76 billion.

If one assumes that there is virtually no chance of a major disruption the Teisberg Model estimates that a change in the BAU price of oil, from \$52.00 per barrel to \$90.00 per barrel will increase the net economic benefit from \$9.11 billion to \$34.01 billion. An increase in the annual probability of a major interruption to 0.10 causes the Teisberg Model to estimates that a change in the price of crude oil from \$52.00 per barrel to \$90.00 per barrel will increase the net economic benefit from \$22.72 billion to \$156.54 billion.

Figure B-2 is the normal probability plot of the effects which appear in Table B-1 and represented by Figure B-1. If the fifteen effects from the model were not due to changes of the input factors then the effects are due to some random variation which is assumed to be normal. If this is the case the normal probability plot of effects should appear more or less as a straight line. Figure B-2 suggests that effects 3, 4, 10, 1, and possibly 6 and 7 are not on the same "straight" line formed by the remaining effects. This plot tends to confirm the identification of significant effects by the method outlined in paragraph A4.

The Teisberg Model
Average Net Economic Benefit

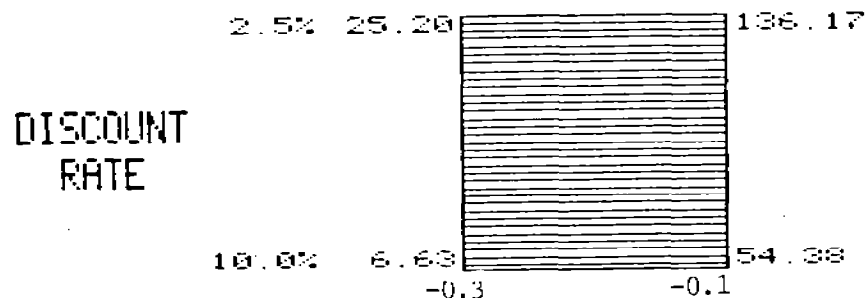
<u>Mean</u>	<u>Estimate</u>
Main Effects	55.59
1. Discount rate	50.18*
2. BAU crude price	21.52
3. Demand elasticities	79.36*
4. Probability of a major disruption	68.07*
2nd Order Interactions	
5. Discount rate X BAU crude price	9.39
6. Discount rate X Demand elasticities	31.61*
7. Discount rate X Probability of a major disruption	28.08*
8. BAU crude price X Demand elasticities	16.25
9. BAU crude price X Probability of a major disruption	21.87
10. Demand elasticities X Probability of a major disruption	54.46*
3rd Order Interactions	
11. Discount rate X BAU crude price X Demand elasticities	5.95
12. Discount rate X BAU crude price X Probability major disruption	8.57
13. Discount rate X Demand elasticities X Probability of a major disruption	21.69
14. BAU crude price X Demand elasticities X Probability of a major disruption	16.66
4th Order Interaction	
15. Discount rate X BAU crude price X Demand elasticities X Probability of a major disruption	6.10
Estimated standard error	13.37
Level of statistical significance at $p < 0.10$	26.95

* Significant effects at $p < 0.10$
Table B-1

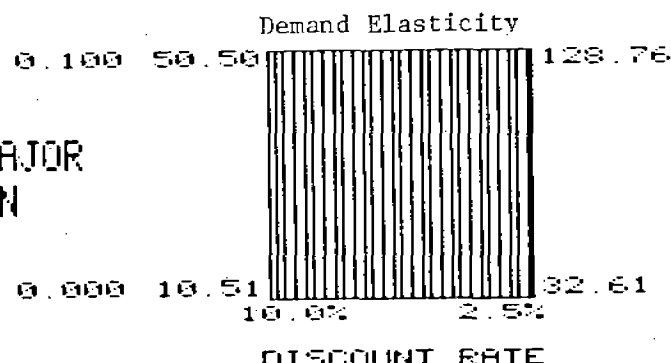
THE TEISBERG MODEL

figure B-1

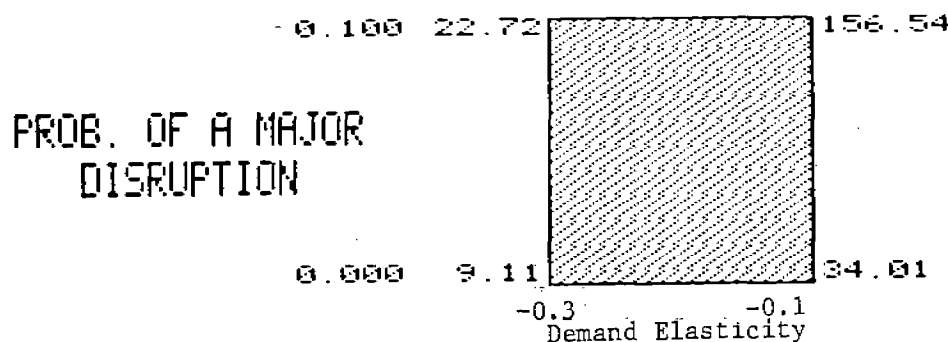
Total U.S. Consumption of Products in MMB0
2nd Order Interactions



The net economic benefit as a function of the discount rate and the Demand Elasticity



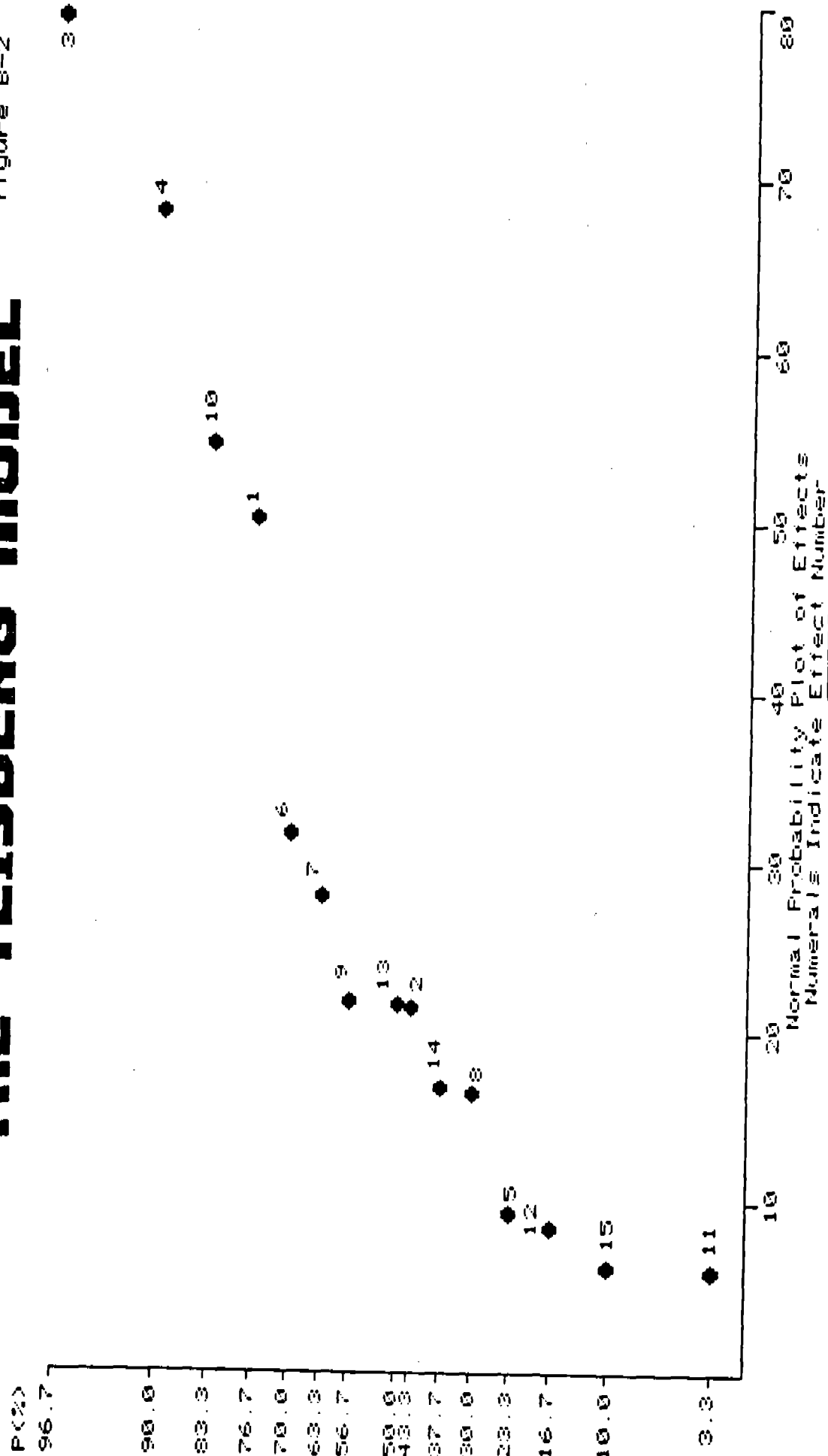
The net economic benefit as a function of the probability of a major disruption and the discount rate.



The net economic benefit as a function of the probability of a major disruption and the Demand Elasticity

THE TEISBERG MODEL

Figure B-2



B2 The Binary Estimate

$$\begin{aligned}\text{Define: } x_d &= \begin{cases} -1 & \text{if } d = 10.0\% \\ 1 & \text{if } d = 2.5\% \end{cases} \\ x_e &= \begin{cases} -1 & \text{if } e = -0.3 \\ 1 & \text{if } e = -0.1 \end{cases} \\ x_p &= \begin{cases} -1 & \text{if } p = 0.000 \\ 1 & \text{if } p = 0.100 \end{cases}\end{aligned}$$

Where d is the discount rate, e is the elasticity of demand, and p is the probability of a major oil disruption.

With the definitions above and the information contained within the analysis of the net economic benefit (section B1) one can construct the following binary estimate:

$$Y = 55.59 + (50.18)/2 x_d + (79.36)/2 x_e + (68.07)/2 x_p + \\ (31.61)/2 x_d x_e + (28.08)/2 x_d x_p + (54.46)/2 x_e x_p$$

or

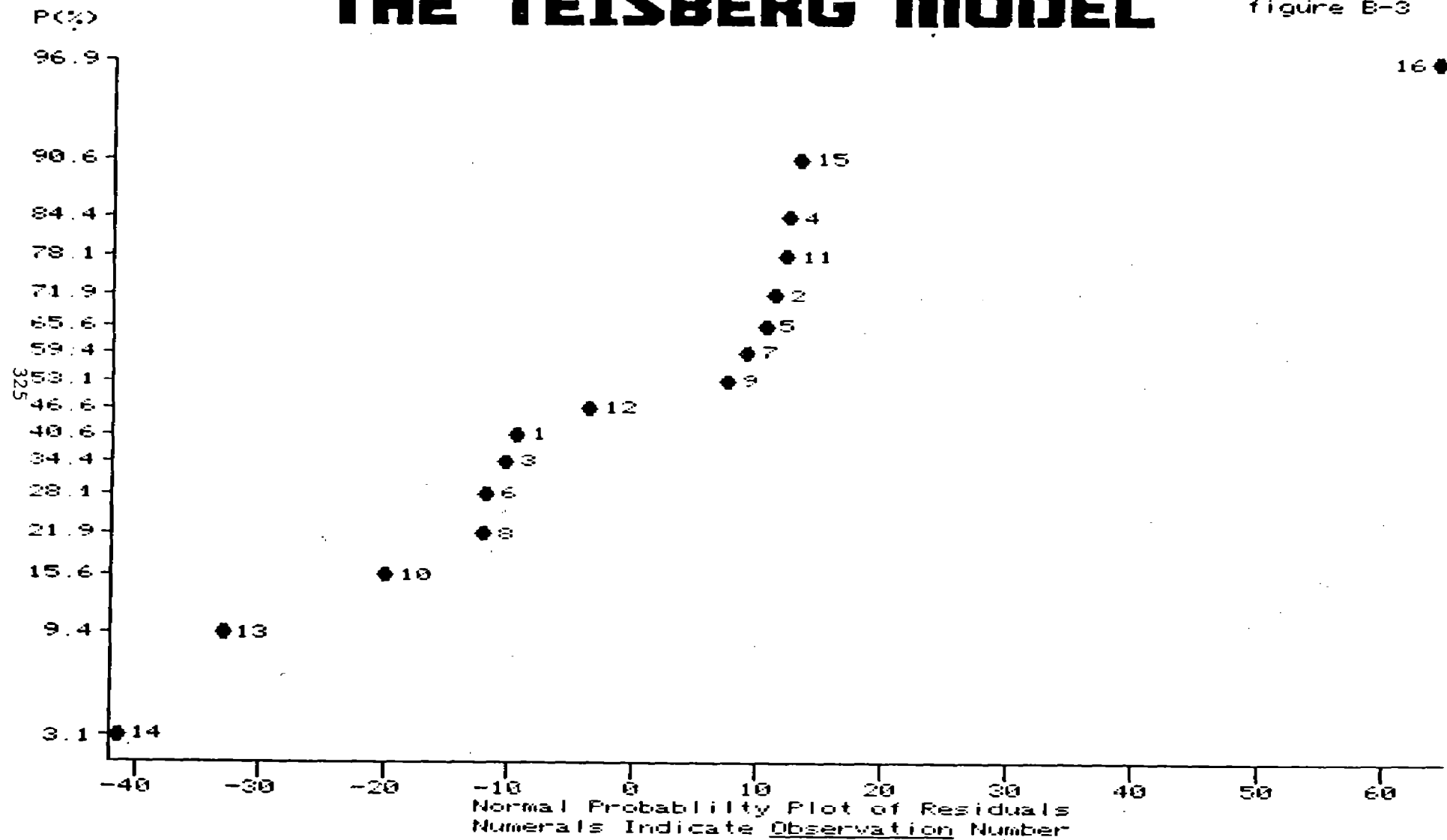
$$Y = 55.59 + 25.09 x_d + 39.68 x_e + 34.04 x_p + \\ 15.81 x_d x_e + 14.04 x_d x_p + 26.23 x_e x_p$$

A normal probability plot of the residuals, figure B-3 can be used to examine the adequacy of this estimate of the Teisberg Model. The residuals for this estimate, are found in Table 1. If all of the variation is explained by the proposed estimating equation then the normal probability plot of residuals will lie more or less on a straight line. Clearly the residual from observation 16 and most likely observations 14 and 13 do not lie on the "straight" line formed by the remaining observations. This suggests that although an improvement in the original estimate has been accomplished, more work remains to be done. Promising avenues of investigation include transforming the data before the application of a factorial design as proposed by Daniel and/or the use of response surface analysis.

THE TEISBERG MODEL

figure B-3

16



B3 The Continuous Estimate

To construct the continuous estimate from the binary estimate replace:

$$X_d \text{ with } \frac{2d - 0.025 - 0.1}{0.025 - 0.1} = \frac{2d - 0.125}{-0.075}$$

$$X_e \text{ with } \frac{2e + 0.1 + 0.3}{-0.1 + 0.3} = \frac{2e + 0.4}{0.2}$$

$$X_p \text{ with } \frac{2p - 0.1}{0.1 - 0} = \frac{2p - 0.1}{0.1}$$

to obtain:

$$\begin{aligned} Y = & 55.59 + 25.098 ((2d - 0.125)/-0.075) \\ & + 39.68 ((2e + 0.4)/0.2) + 34.04 ((2p - 0.1)/0.1) \\ & + 15.81 ((2d - 0.125)/-0.075)((2e + 0.4)/0.2) \\ & + 14.04 ((2d - 0.125)/-0.075)((2p - 0.1)/0.1) \\ & + 26.23 ((2e + 0.4)/0.2)((2p - 0.1)/0.1) \end{aligned}$$

which simplifies to:

$$\begin{aligned} Y = & 119.57 - 1,137.87 d + 398.00 e + 2198.00 p \\ & - 4216.00 de - 7488.00 dp + 5246.00 ep \end{aligned}$$

B4 The Differential Estimate

If $c(w)$ denotes the change in the variable w , then the estimate of the change of the net benefit is:

$$\begin{aligned} c(y) = & -1137.87 c(d) + 398.00 c(e) + 2198.00 c(p) \\ & - 4216.00 d c(e) - 4216.00 c(d) e \\ & - 7888.00 d c(p) - 788.00 c(d) p \\ & + 5246.00 e c(p) + 5246.00 c(e) p \end{aligned}$$

Although this was developed as a global estimate it can be used for local approximations. If the model has been evaluated for a set of input factors (d, e, p) and one wishes to estimate the net economic benefit for a point (d', e', p') which is close to (d, e, p) then calculate the $c(y)$, the change in the net economic benefit and add that value to the model's estimate for the point (d, e, p) .

Bibliography

- Box, G.P.E. and D.R. Cox (1964). An analysis of transformations, J. Roy. Stat. Soc., Series B, 26,211.
- Box, G.P.E. and N.R. Draper (1969). Evolutionary Operation: A Statistical Method for Process Improvement, Wiley.
- Box, G.P.E. and J.S. Hunter (1961) The 2^k -P factorial design, Technometrics, 3,311,449.
- Box, G.P.E., J.S. Hunter and W.G. Hunter (1978) Statistics for Experimenters, Wiley.
- Chilman, W.J. (1983) Economic Analysis of Four Alternative SPR Development Options Employing the Teisberg Model in a Dynamic Programming Treatment of Microeconomic Factors Associated with Costs and Benefits Reduced to Net Present Value Bases, Dept. of Energy May 83.
- Churchill, R.V. (1960). Complex Variables and Applications, McGraw-Hill.
- Cochran, W.G. and G.M. Cox (1957). Experimental Design, 2nd ed. Wiley.
- Daniel, C. (1959). Use of half - normal plot interpreting factorial two-level experiments, Technometrics, 1,149.
- Daniel, C. (1976). Applications of Statistics to Industrial Experimentation, Wiley.
- Davies, O.L. (Ed.) (1971). The Design and Analysis of Industrial Experiments, Hafner (Macmillan).
- Hunter, J.S. (1966). Inverse Yates algorithm, Technometrics.
- Hunter, W.G. (1976). Some ideas about teaching design of experiments, with 2⁵ examples of experiments conducted by students, Am. Stat., 31,12.
- Hunter, W.G. and E. Chacko (1971). Increasing industrial productivity in developing countries, Int. Rev. Div., 13,3,11.
- Jenkins G.M. (1969). A systems study of a petrochemical plant J. Syst.Enh., 1, 90.
- Maske, D.M. (1970). High - rate, fine-mesh screening of combined wastewater flows, J. Water Pollut. Control Fed., 42, 1476.
- Plackett, R.L. and J. P. Burman (1946). The design of optimum multifactorial experiments, Biometrics, 33,305.
- Tessmer, J.M. (1982). Validation of the International Petroleum Model for the SPR Crude Mix Study, Dept. of Energy Spring 1982.
- Tessmer, J.M. (1982). An example of Software Validation using a Factorial Design, ARO Report 83-2, Proceedings of the Twenty Eight Conference on the Design of Experiments, Monterey California, Oct. 82.
- Yates, F. (1937). The Design and Analysis of Factorial Experiments, Bulletin 35, Imperial Bureau of Soil Science, Harpenden, Herts England, Hafner (Macmillan).
- Yates F. (1970). Selected Papers, Hafner (Macmillan).
- Youden, W.J. (1972). Enduring values, Technometrics, 14,1.

HIGH TO LOW DOSE EXTRAPOLATION OF EXPERIMENTAL ANIMAL CARCINOGENESIS STUDIES

Charles C. Brown

National Cancer Institute

Bethesda, MD 20205

ABSTRACT

Quantitative risk assessment requires extrapolation from results of experimental assays conducted at high dose levels to predicted effects at lower dose levels which correspond to human exposures. The meaning of this high to low dose extrapolation within an animal species will be discussed, along with its inherent limitations. A number of commonly used mathematical models of dose-response necessary for this extrapolation, will be discussed. Other limitations in their ability to provide precise quantitative low dose risk estimates will also be discussed. These include: the existence of thresholds; incorporation of background, or spontaneous responses; modification of the dose-response by pharmacokinetic processes.

In recent years, as the serious long-range health hazards of environmental carcinogens have become recognized, the need has arisen to quantitatively estimate the effects upon humans exposed to low levels of these agents. Inherent in this estimation procedure is the necessity to extrapolate evidence observed under one set of conditions in one population group or biological system to arrive at an estimate of the effects expected in the population of interest under another set of conditions.

The quantitative assessment of human health risk from exposure to carcinogenic agents is often approached by relating the exposure level of the agent to a measure of the cancer risk as determined from experimental data on animals or other biological systems. For the extrapolation of animal study results to man, much care should be placed in the design and conduct of these studies, since many factors may influence their results. These factors include the dosage and frequency of exposure, route of administration, species, strain, sex and age of the animal, duration of the study, and various other modifying factors as deemed important for the particular agent and effect being studied.

Experimental animal bioassays to measure the dose-response of the agent in question must necessarily be based on exposure levels higher than those for which the risk estimation is to be made. A limited number of experimental animals requires high exposure levels in order to measure a carcinogenic effect if it exists. Some consideration has been given to the possibility of conducting extremely large experiments at very low dose levels. However, as Schneiderman, et al. (1) remark, "purely logistical problems might guarantee failure." Therefore, to obtain reliably measureable effects, the experimental information must be based on levels of exposure high enough to detect positive results. Since large segments of the human populations are often exposed to

much lower levels, these data at high exposure levels must be extrapolated to lower levels which correspond to human exposure. The purpose of this presentation is to describe the current statistical methods used for this "high to low dose" extrapolation in experimental animal species and to emphasize the uncertainties necessarily attached to the estimates made with these methodologies.

The high to low dose extrapolation problem is conceptually straightforward. Since the risk at low exposure levels cannot be measured by direct experimentation, an assumed mathematical relationship between dose (exposure) and response (risk) must be used to extrapolate from the high experimental doses to the low environmental levels. The probability of a toxic response is modeled by a dose-response function $P(D)$ which represents the probability of a carcinogenic response when exposed to D units of the carcinogenic agent. A general mathematical model is chosen to describe this functional relationship, its unknown parameters are estimated from the available data, and this estimated dose-response function $P(D)$ is then used to either: (i) estimate the response measure at a particular low dose level of interest; or (ii) estimate that dose level corresponding to a desired low level of response (this dose estimate is commonly known as the virtually safe dose, VSD).

Many mathematical dose-response models have been proposed for this problem. The following section describes the more commonly used models.

Mathematical Models of Dose-Response

To estimate the effects expected outside the range of observable experimental data, a mathematical model relating dose, i.e., level of exposure to the toxic agent, to response, i.e., a quantitative measure of the deleterious effect produced, is necessary. In general terms, dose-response is

the relation between a measureable stimulus, physical, chemical or biological, and the response of living matter measured in terms of the reaction produced over some range of the degree or level of the stimulus.

Tolerance Distribution Models

When the response is quantal (whether or not a specific effect is produced), its occurrence for any particular subject will depend upon the level of the stimulus. For this subject under constant environmental conditions, a common assumption is that there is a certain dose level below which the particular subject will not respond in a specified manner, and above which the subject will respond with certainty. This level is referred to as the subject's tolerance. Because of biological variability among subjects in the population, their tolerance levels will also vary. For quantal responses, it is therefore natural to consider the frequency distribution of tolerances over the population studied. If D represents the level of a particular stimulus, or dose, then the frequency distribution of tolerances, $f(D)$, may be mathematically expressed as

$$f(D) = dP(D)/dD,$$

which represents the proportion of subjects whose tolerances lie between D and $D+dD$, where dD is small. If all subjects in the population are exposed to a dose of D_0 , then all subjects with tolerances less than or equal to D_0 will respond, and the proportion, $P(D_0)$, this represents of the total population is given by

$$P(D_0) = \int_0^{D_0} f(D) dD .$$

Assuming that all subjects in the population will respond to a sufficiently high dose level, then

$$P(\infty) = \int_0^{\infty} f(D)dD = 1 .$$

The function $P(D)$ can be thought of as representing the dose-response either for the population as a whole, or for a subject randomly selected from the population. The notion that a tolerance distribution, or dose-response function, could be determined solely from consideration of the statistical characteristics of a study population was introduced independently by Gaddum (2) and Bliss (3).

The results of toxicity tests have often shown that the proportion of responders increases monotonically with dose and often exhibits a sigmoid relationship with the logarithm of the exposure level. This observation led to the development of the log normal, or probit, model for the tolerance frequency distribution,

$$f(D; \mu, \sigma) = (2\pi\sigma^2)^{-1/2} \exp - \frac{1}{2} \left(\frac{\log(D) - \mu}{\sigma} \right)^2, \quad \sigma > 0,$$

while the dose-response function is given by the cumulative normal probability,

$$P(D; \mu, \sigma) = \Phi[(\log(D) - \mu)/\sigma] .$$

where μ and σ^2 represent the mean and variance of the distribution of the log tolerances. This method was put into its modern form by Bliss (4), and Finney (5) gives a brief history of its development.

Other mathematical models of tolerance distributions which produce a sigmoid appearance of their corresponding dose-response functions have been suggested. The most commonly used is the log logistic function,

$$P(D; a, b) = [1 + \exp(a + b \log(D))]^{-1}, \quad b < 0,$$

which, like the log normal model is sigmoid and symmetric about the 50% response level, but approaches the extremes, 0% and 100% response, more slowly

than does the log normal. The logistic function has been derived from chemical kinetic theory, and was proposed as a dose-response model by Worcester and Wilson (6) and Berkson (7). The log logistic and log normal functions are similar in appearance so that discrimination between them is nearly impossible.

Models Derived From Mechanistic Assumptions

A number of dose-response models have been suggested on the basis of assumptions regarding the mechanism of action of the toxic agent upon its target site. The "hit theory" for interaction between radiation particles and susceptible biologic targets has generated a general class of these models (8). This theory is also applicable to the action of chemical toxicants upon their target sites. In general, this theory rests upon a number of postulates, which include: (1) the organism has some number M of "critical targets" (usually assumed to be infinitely large); (2) the organism responds if m or more of these critical targets are "destroyed"; (3) a critical target is destroyed if it is "hit" by k or more toxic particles; and (4) the probability of a hit in the low dose region is proportional to the dose level of the toxic agent, i.e. $\text{Prob}(\text{hit}) = \lambda D$, $\lambda > 0$.

Some commonly used special cases of this general theory are the single-hit model,

$$P(D; \lambda) = 1 - \exp(-\lambda D) ,$$

where the subject responds if a single critical target is destroyed by a single hit; and the multihit model,

$$P(D; \lambda, k) = \int_0^{\lambda D} \frac{x^{k-1} \exp(-x) dx}{\Gamma(k)} ,$$

where $\Gamma(k)$ denotes the gamma function, and the subject responds if a single

critical target is destroyed by k hits. For a discussion of the single-hit model as applied to the high to low dose extrapolation problem, see (9,10); the Report of the Scientific Committee of the Food Safety Council (11,12) and Rai and Van Ryzin (13,14) discuss the application of the multihit model for dose extrapolation.

Other mechanistic models have also been derived from quantitative theories of carcinogenesis. The multistage carcinogenesis theory (15-17) assumes that a single cell can generate a malignant tumor only after it has undergone a certain number, k , of heritable changes. This theory leads to the multistage model,

$$P(D; \lambda_1, \dots, \lambda_k) = 1 - \exp(-(\lambda_1 D + \lambda_2 D^2 + \dots + \lambda_k D^k)), \quad \lambda_i \geq 0 \quad i=1, \dots, k.$$

The use of this model for extrapolation purposes has been described by Brown (18) and Guess and Crump (19,20).

The multicell carcinogenesis theory of Fisher and Holloman (21) leads to a dose-response function having extrapolation characteristics similar to the multihit model,

$$P(D; \lambda, k) = 1 - \exp(-\lambda D^k), \quad \lambda, k > 0.$$

This model has also been termed the Weibull model and Van Ryzin (22) discusses its application to extrapolation problems.

Discrimination among dose-response models

Given a postulated functional form of the dose-response relationship, the experimental data is used to estimate the unknown parameters. It might be thought that the basis for selection of one particular model over the others would be provided by the observed dose-response. However, this is often not the case, as many dose-response models appear similar to one another over the

range of observable response rates. Figures 1 and 2 compare the dose-response relationships of the more commonly used models; Figure 1 compares the log normal, log logistic and single-hit models; Figure 2 compares the multihit, Weibull and multistage models.

In the left panel of Figure 1, the parameters for these models were chosen to make the response rates equal at dose levels of 1 and $1/4$; in the left panel of Figure 2, the parameters for the models were chosen to make the response rates equal at dose levels of 2 and 0.5. These figures clearly show that it would take an inordinately large set of experimental or observational data to be able to conclude which of the models provide a significantly better fit to an observed dose-response.

If the estimated dose-response is to be used to predict the response rate that would be expected from an exposure level within the range of observable rates, then the models within each of the two sets compared will give similar results. However, extrapolation to exposure levels expected to give very low response rates is highly dependent upon the choice of model, as shown in the right panels of Figures 1 and 2. These figures extend the dose-response in the left panels to much lower dose levels. The further one extrapolates from the observable response range, the more divergent the models become. At a dose level which is $1/1000$ of the dose giving a 50% response, the single-hit model gives an estimated response rate 200 times that of the lognormal model, and the multistage model gives an estimated response rate over 210 times that of the multihit model.

Krewski and Van Ryzin (23) examined the extrapolation characteristics of these six commonly used dose-response models. They applied these models to 20 sets of toxic response data that were taken from the Report of the Scientific Committee of the Food Safety Council (11,12). The toxic responses were both

carcinogenic and noncarcinogenic in nature. Of the 19 data sets showing an convex (i.e. upward curvature) dose-response, all estimates of the virtually safe dose (VSD) at a response rate of $P = 10^{-5}$ or smaller had the ordering, single-hit < multistage < (Weibull, log logistic, multihit) < log normal. That is, the Weibull, log logistic, and multihit produce VSD's of approximately the same order of magnitude, the single-hit model produces the smallest VSD, and the log normal model the largest VSD. In addition, the difference between the extremes, the single-hit and log normal models, is often several orders of magnitude.

Table I and Figure 3 give an example of this behavior for these models applied to the incidence of liver hepatomas in mice exposed to various levels of DDT (24). Table I shows that each of the six dose-response models fit the observed data nearly equally well (the multistage model fits the data as well as the others). Therefore, the data in the observable response range (for this study, between 2 and 250 ppm DDT in the daily diet) cannot discriminate among these models. Based on the goodness-of-fit statistics, the Weibull model fits the best ($P = 0.22$), but not significantly better than any of the other models. However, there is a significant difference among the VSD estimated from these models; the log normal model estimates a VSD over 3000 times as large as the single-hit model. Therefore, these experimental data leave the true VSD open to wide speculation.

The fact that an experimental study conducted at exposure levels high enough to give measureable response rates cannot clearly discriminate among these various models, along with the fact that those models show substantial divergence at low exposure levels present one of the major difficulties for the problem of low dose extrapolation. Since the multistage model has the extrapolation characteristics of most other models, Brown (25) has suggested

its use to provide estimates of both sampling and model variability for this low dose extrapolation problem.

Adjustments for Natural Responsiveness

The mathematical dose-response models described in the preceding sections have assumed responses of the subjects to be due solely to the applied stimuli. However, many toxicity experiments and observational studies show clear evidence that responses can occur even at a zero dose. Thus, any mathematical dose-response function should properly allow for this natural, or 'background', responsiveness.

Two methods have been proposed to incorporate the possibility of response due to factors other than the stimulus in question. The first is commonly known as 'Abbott's correction' which is based on the assumption of an independent action between the stimulus and the background (26). If the probability of response in the absence of any stimulus is denoted by P_0 , then the overall response probability at dose level D , assuming independent actions, becomes

$$P(D) = P_0 + (1-P_0)P^*(D) ,$$

where $P^*(D)$ represents the dose-induced probability of response. The second method assumes that the dose acts in an additive manner with the background environment, producing the overall dose-response model (27)

$$P(D) = P^*(D+D_0) ,$$

where D_0 represents some unknown background level of the stimulus (or other stimuli that produce the response in a mechanistically dose-additive manner).

It is often difficult to discriminate between the independent and additivity assumption on the basis of dose-response data. Figure 4 compares the theoretical dose-response relationships of these two assumptions where

$P^*(D)$ is a log logistic model. The parameters of these models were chosen to minimize their difference. Clearly, a large set of data would be required to determine the proper manner to incorporate background response. To describe the dose-response in this observable response range, this figure shows that this assumption is not an important issue, as both will describe equally data in the observable response range. However, for purposes of low-dose extrapolation, this assumption can have important consequences. Crump, et al. (17) have shown mathematically, that no matter what dose-response model, $P^*(D)$, is used, the additivity assumption will lead to a linear dose-response in the low dose region. This will not be true for the independent action assumption. Hoel (28) compares low dose risk extrapolations based on the two assumptions applied to a log normal dose-response model. His results are given in Table II. This table clearly shows the low-dose linearity of the additive assumption, and the substantial difference between the additive and independence assumptions at low dose levels. Hoel also examined models which incorporate a mixture of independent and additive background response, and found that low dose linearity prevails except when the background mechanism is totally independent of the dose-induced mechanism.

Pharmacokinetic Models

Pharmacokinetic hypotheses concerning toxicity from foreign chemicals state that biological effects are manifestations of biochemical interactions between the foreign substances (or substances derived from them) and components of the body. A critical problem in the application of pharmacokinetic principles to risk extrapolation is the potential change in metabolism or other biochemical reactions as external exposure levels of the toxic agent decrease. Linear pharmacokinetic models are often used. However,

there are numerous examples of nonlinear behavior in the dose range studied, and these nonlinear kinetics pose significant problems for quantitative extrapolation from "high" to "low" doses if the kinetic parameters are not measured (29-31).

As shown in Figure 5, linear kinetics assume that the reaction rate per unit time of a chemical reaction is proportional to the concentration C of the substance being acted upon; whereas nonlinear kinetics are most often described in the form of a Michaelis-Menten expression, often referred to as "saturable" kinetics. If all processes are linear, then the concentration rate of the toxic substance at its site of action ('effective dose') will be proportional to the external exposure rate ('administered dose'). However, saturation phenomena may produce different results depending upon the processes affected; if elimination and/or detoxification pathways are saturable, then the effective dose will increase more rapidly with the administered dose than linear kinetics would suggest; if the distribution and/or activation pathways are saturable, then the effective dose will increase less rapidly with the administered dose.

Gehring and Blau (32) and Gehring, et al. (33) discuss pharmacokinetic models with respect to extrapolation of carcinogenic risk from high to low doses. As an example, Gehring et al. (29) applied pharmacokinetic principles to the dose-response of hepatic angiosarcomas in rats exposed to different concentrations of atmospheric vinyl chloride over a period of 12 months. The results of their study are shown in Figure 6. Since the metabolic activation of vinyl chloride appears to be a saturable process, the observed relationship between response, as measured by the proportion of rats with hepatic angiosarcomas, and dose, as measured by the external atmospheric exposure level of vinyl chloride, is clearly nonlinear, showing a leveling out at the

highest exposure levels which cannot be explained by a number of the previously discussed dose-response models (e.g. log normal and multistage). However, if dose is measured in terms of the amount of vinyl chloride metabolized, then the dose-response becomes much more linear, and most models provide an adequate fit to the data.

Summary and Conclusions

The preceeding sections have discussed the general problem of high dose to low dose extrapolation. The purpose of this extrapolation is to estimate the effects of low level exposure to carcinogenic agents known to be associated with undesired effects at high dose levels.

Mathematical models of dose-response are necessary for this extrapolation process since the low dose effects, expected to be on the order of response rates of 10^{-6} , are too small to be accurately measured with limited study sample sizes. A number of mathematical dose-response models have been proposed for extrapolation purposes; we previously saw how similar they can appear to one another in the range of observable response rates, yet how different they become at lower, unobservable response rates, the region of primary interest. This is the single, most important limitation of this extrapolation methodology. An estimate of risk at a particular low dose, or an estimate of the dose leading to a prespecific level of risk is highly dependent upon the mathematical form of the presumed dose-response; we have seen that differences of 3 - 4 orders of magnitude are not uncommon.

Pharmacokinetic information on the fate of a toxic agent once it enters the body is beginning to be incorporated into the high to low dose extrapolation process. Nonlinear kinetics may be an important determinant of the nonlinear dose-response relationships often observed in experimental

studies of toxic agents. As noted in previously, Gehring et al (29) have shown that the metabolism of inhaled vinyl chloride is a saturable process that provides one explanation of the concave liver carcinogenesis dose-response observed in animal studies. In a study of urethane-induced pulmonary adenomas shown in Figure 7, White (30) found that the convex relationship between the amount of urethane injected into a mouse lung and the number of subsequent lung adenomas could be explained by nonlinear kinetics of excretion. Such pharmacokinetic models and dose-response studies of the kinetics of physiological processes might considerably strengthen the ability to extrapolate from high to low dose levels.

Other sources of uncertainty in high to low dose extrapolation include: (1) the possible existence of thresholds; (2) heterogeneity of sensitivity to the toxic agent among members of the exposed population; and (3) mechanisms of action for carcinogens (i.e. whether the agent initiates the process or acts at a later stage). The existence of a single threshold for the entire exposed population should allow for estimation of a clearly safe level of exposure. However, its estimation could be associated with a high degree of uncertainty. Heterogeneity in individual thresholds and sensitivity to the toxic agent induces additional uncertainty in high to low dose extrapolations. The theoretical relationship of dose rate and duration of exposure to cancer risk indicates that similar exposure patterns (i.e. same dose rate and duration) will not necessarily lead to similar levels of risk since the age at exposure may also be an important determinant of risk. Thus, uncertainty in the mechanism of toxic action induces another potentially large uncertainty into risk extrapolations.

Therefore, all these sources of uncertainty, (1) dose-response model, (2) pharmacokinetic behavior of the toxic agent, (3) thresholds, (4) heterogeneity, and (5) mechanisms of action, lead to potentially enormous variation in estimates of risk from high to low dose extrapolations.

Literature Cited.

1. Schneiderman, M.A.; Mantel, N.; Brown, C.C. Ann. NY Acad. Sci. 1975, 246, 237-246.
2. Gaddum, J.H. Medical Research Council, Special Report Series No. 183, 1933.
3. Bliss, C.I. Science 1934, 79, 38-39.
4. Bliss, C.I. Ann. Appl. Biol. 1935, 22, 134-167.
5. Finney, D.J. "Probit Analysis, 3rd Ed." Cambridge University Press: London, 1971.
6. Worcester, J.; Wilson, E.B. Proc. Natl. Acad. Sci. 1943, 29, 78-85.
7. Berkson, J. J. Amer. Stat. Assn. 1944, 39, 134-167.
8. Turner, M. Math. Biosci. 1975, 23, 219-235.
9. Hoel, D.G.; Gaylor, D.; Kirschstein, R.; Saffiotti, U.; Schneiderman, M. J. Tox. and Environ. Health 1975, 1, 133-151.
10. "The Effects of Populations of Exposure to Low Levels of Ionizing Radiation," National Academy of Sciences, 1980.
11. Scientific Committee, Food Safety Council. Food and Cosmetic Tox. 1978, 16 supplement 2, 1-136.
12. Scientific Committee, Food Safety Council "Proposed System for Food Safety Assessment"; Food Safety Council: Washington, D.C., 1980; p. 137.
13. Rai, K.; Van Ryzin, J. in "Energy and Health"; Breslow, N.; Whittemore, A. Eds.; SIAM: Philadelphia, 1979; p. 99.
14. Rai, K. and Van Ryzin, J. Biometrics 1981, 37, 341-352.
15. Armitage, P; Doll, R. Brit J. Ca. 1954, 8 1-12.

16. Armitage, P; Doll, R. in "Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 4)"; Neyman, J. Ed.; University of California Press: Berkely and Los Angeles, 1961, p. 19.
17. Crump, K.S.; Hoel, D.; Langley, C.; Peto, R. Cancer Res. 1976, 36, 2973-2979.
18. Brown, C. J. Natl. Cancer Instit. 1978, 60, 101-108.
19. Guess, H.A.; Crump, K.S. Math. Biosci. 1976, 32, 15-36.
20. Guess, H.A.; Crump, K.S. Environ. Health Perspect. 1978, 22, 149-152.
21. Fisher, J.C.; Holloman, J.H. Cancer 1951, 4, 916-918.
22. Van Ryzin, J. J. Occup. Med. 1980, 22, 321-326.
23. Krewski, D.; Van Ryzin, J. in "Current Topics in Probability and Statistics"; Sorgo, M.; Dowson, D.; Rao, J.N.K.; Saleh, E. Eds.; North-Holland: New York, 1981; p. 201.
24. Tomatis, L.; Turusov, V.; Day, N.; Charles, R.T. Int. J. Cancer 1972, 10, 489-506.
25. Brown, C.C. Environ. Health Perspect. 1978, 22, 183-184.
26. Abbott, W.S. J. Econ. Ent. 1925, 18, 265-267.
27. Albert, R; Altshuler, B. in "Radionuclide Carcinogenesis"; Ballou, J.; Mahlum, D.; Sanders, C. Eds.; AEC Symposium Series, Conf-720505, 1973, p. 233.
28. Hoel, D.G. Fed. Proceed. 1980, 39, 67-69.
29. Gehring, P.J.; Watanabe, P.G.; Park, C.N. Tox. App. Pharm. 1978, 44, 581-591.
30. White, M. in "Proceedings of the Sixth Berkely Symposium on Mathematical Statistics and Probability, Vol. 4"; Lecam, L.E.; Neyman, J.; Scott, E.L. Eds.; University of California Press: Berkeley and Los Angeles, 1972; p. 287.

31. Hoel, D.G.; Kaplan, N.L.; Anderson, M.W. Science 1983, 219, 1032-1037.
32. Gehring, P.J.; Blau, G.E. J. Envir. Path and Tox. 1977, 1, 163-179.
33. Gehring, P.J.; Watanabe, P.G.; Young, J.D. in "Origins of Human Cancer (Book A: Incidence of Cancer in Humans)"; Hiah, H.H; Watson, J.D.; Winston, J.A. Eds.; Spring Harbour Laboratory: Cold Springs Harbour, 1977; p. 187.
34. Altshuler, B. in "Environmental Health, Quantitative Methods"; Whittemore, A. Ed.; Society for Industrial and Applied Mathematics, Philadelphia, 1977; p. 31.

Table I: Comparison of Virtually Safe Doses (VSD)
 Leading to an Excess Risk of 10^{-6}
 for Various Dose-Response Extrapolation Models
 models applied to data from (24)

Extrapolation Model	VSD* (ppm DDT in daily diet)	Goodness-of-fit Statistic of Model to Observed Data		
		χ^2	(d.f.)	P-value
Log normal	6.8×10^{-1}	3.93	(2)	0.14
Weibull	5.0×10^{-2}	3.01	(2)	0.22
Multihit	1.3×10^{-2}	3.31	(2)	0.19
Log logistic	6.6×10^{-3}	3.45	(2)	0.18
Multistage	2.5×10^{-4}	-----**		
Single-hit	2.1×10^{-4}	5.10	(3)	0.16

* 97.5% lower confidence limit on VSD

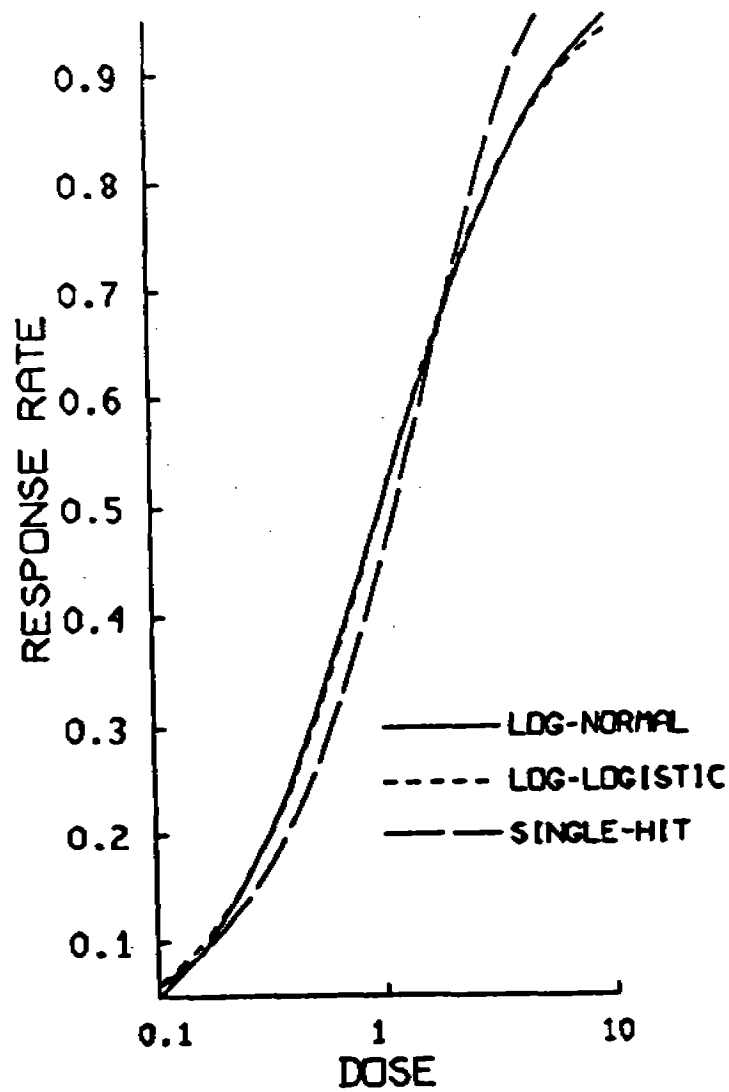
** no goodness-of-fit statistic since the number of parameters
 equals the number of data points

Table 11: Excess Risk $P(D)-P(0)$ for Log Normal Dose Response
Model Assuming Independent and Additive Background

Dose (D)	Type of Background	
	Independent	Additive
10^0	4.0×10^{-1}	4.0×10^{-1}
10^{-1}	1.5×10^{-2}	5.2×10^{-2}
10^{-2}	1.6×10^{-5}	5.2×10^{-3}
10^{-3}	3.8×10^{-10}	5.1×10^{-4}
10^{-4}	1.8×10^{-16}	5.1×10^{-5}

* $P(0) = 0.1$; log normal model slope = 2 from (28)

OBSERVABLE RESPONSE RANGE



UNOBSERVABLE RESPONSE RANGE

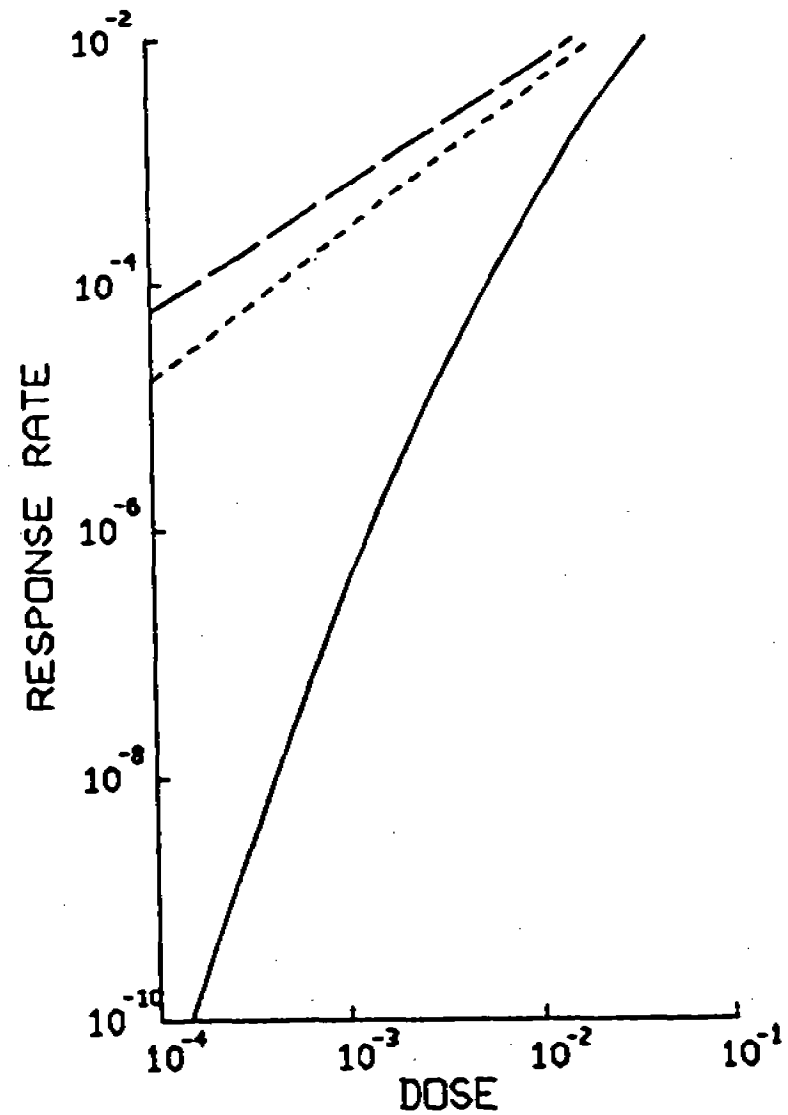


FIGURE 1: Comparison of log-normal, log-logistic and single-hit dose response models

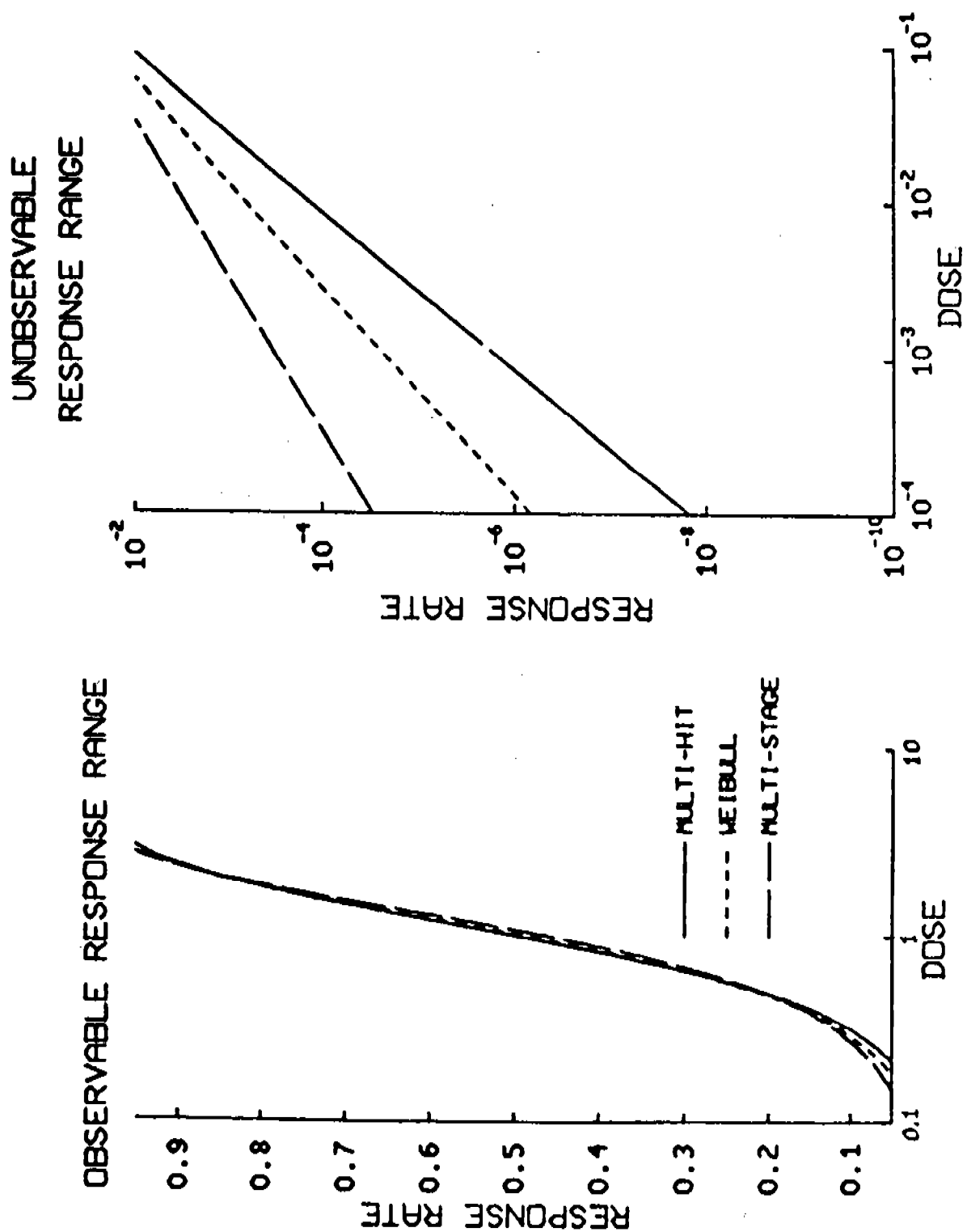
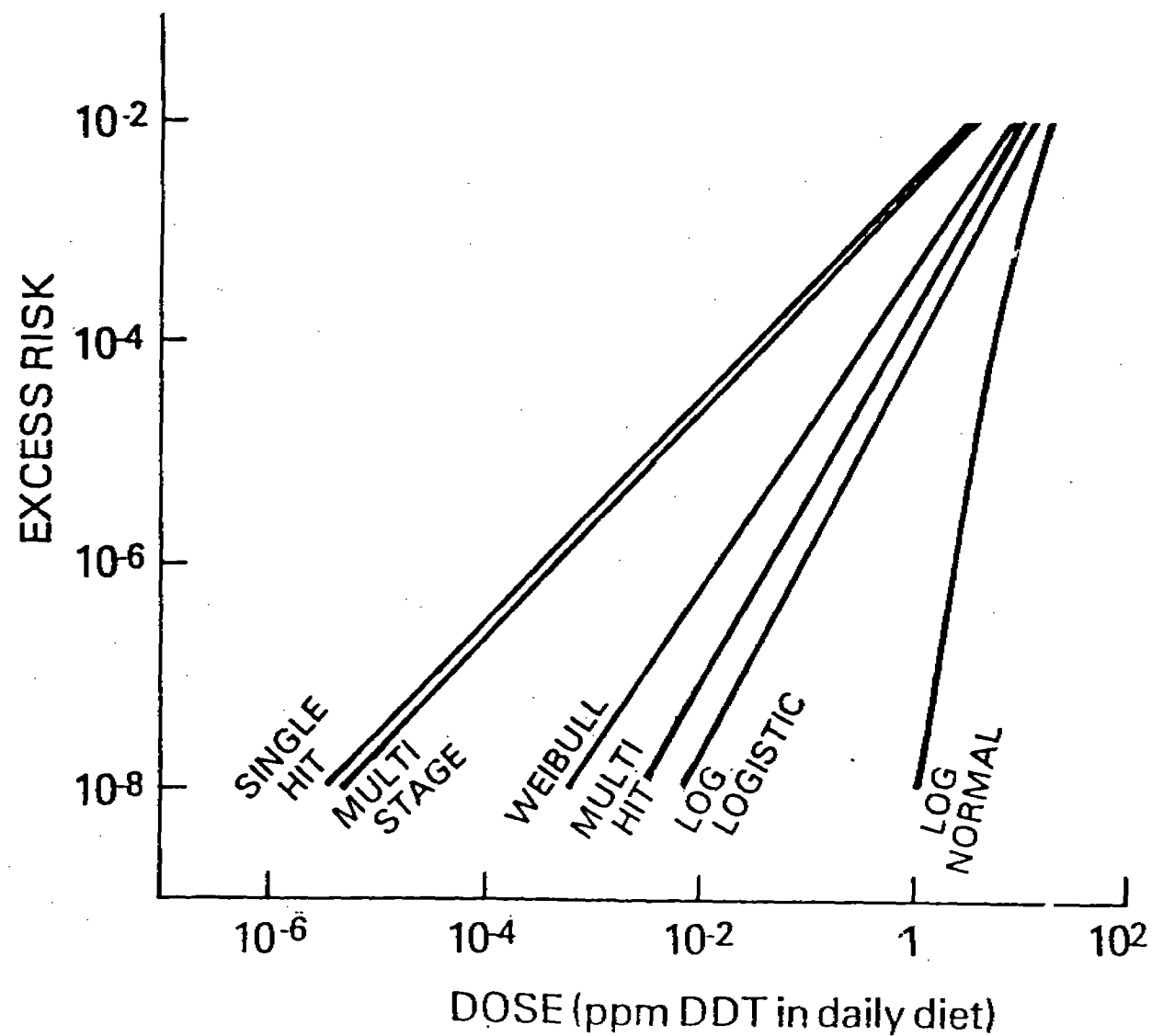


FIGURE 2: Comparison of multi-hit, Weibull and multi-stage dose response models

FIGURE 3: Comparison of high to low dose extrapolations from 6 dose-response models [data from Tomatis et al. (24)]



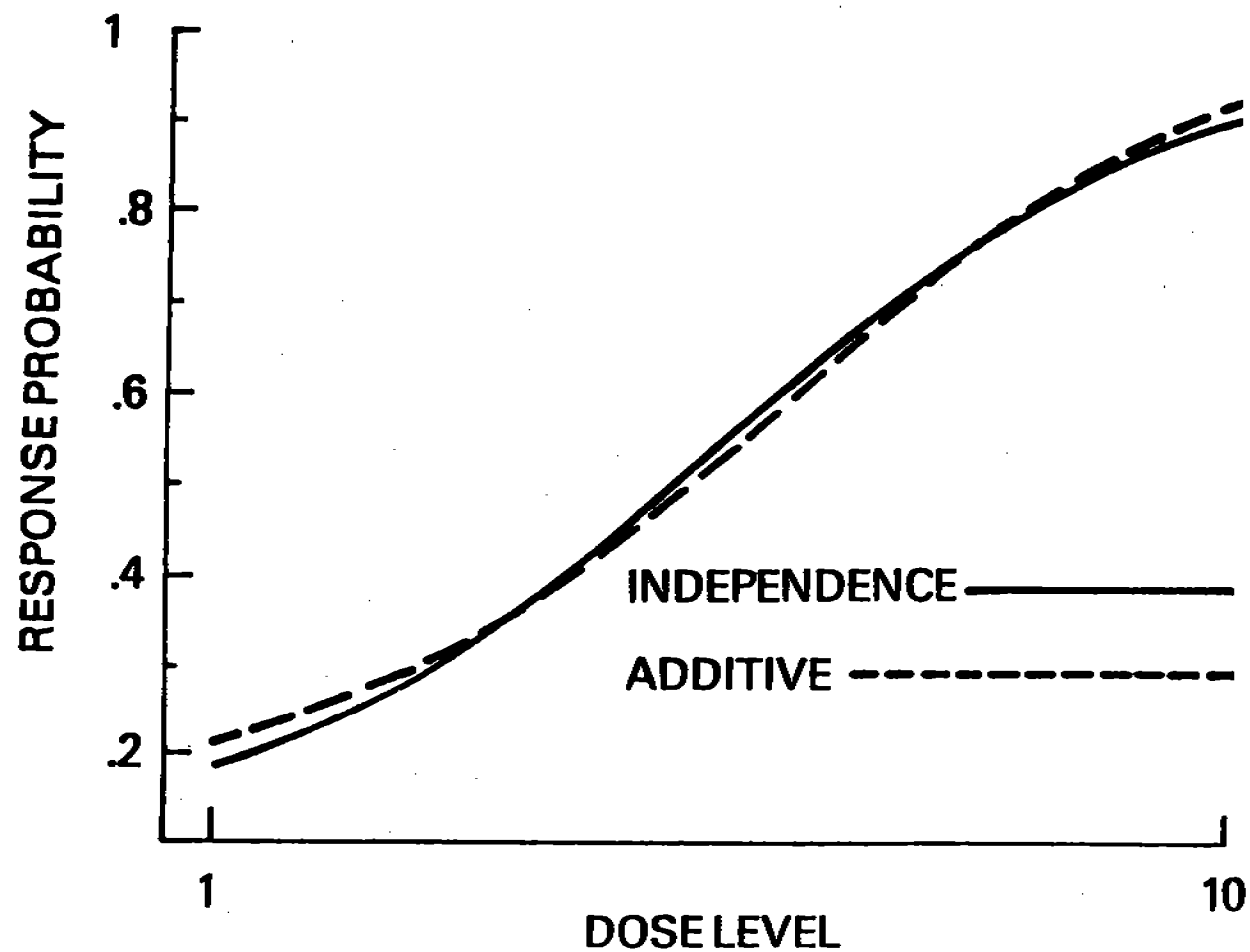
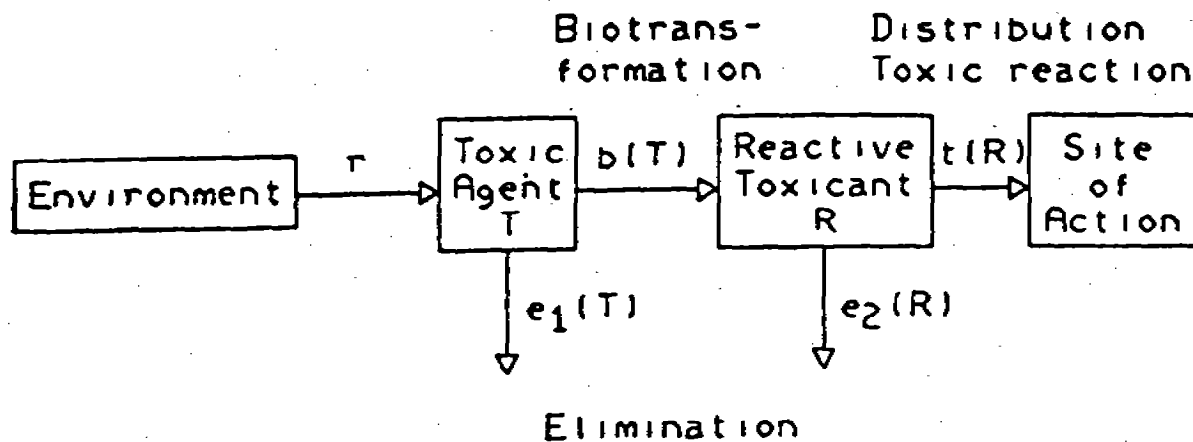


FIGURE 4: Comparison of log logistic dose-response models assuming independent and additive background



REACTION KINETICS

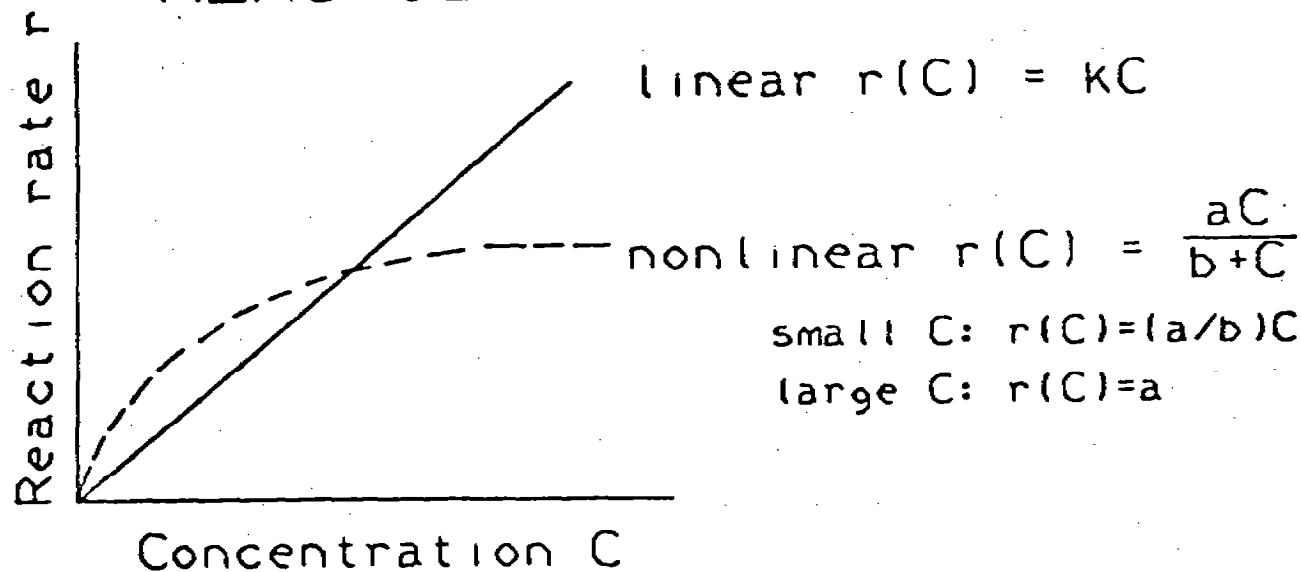


FIGURE 5: Description of general pharmacokinetic model

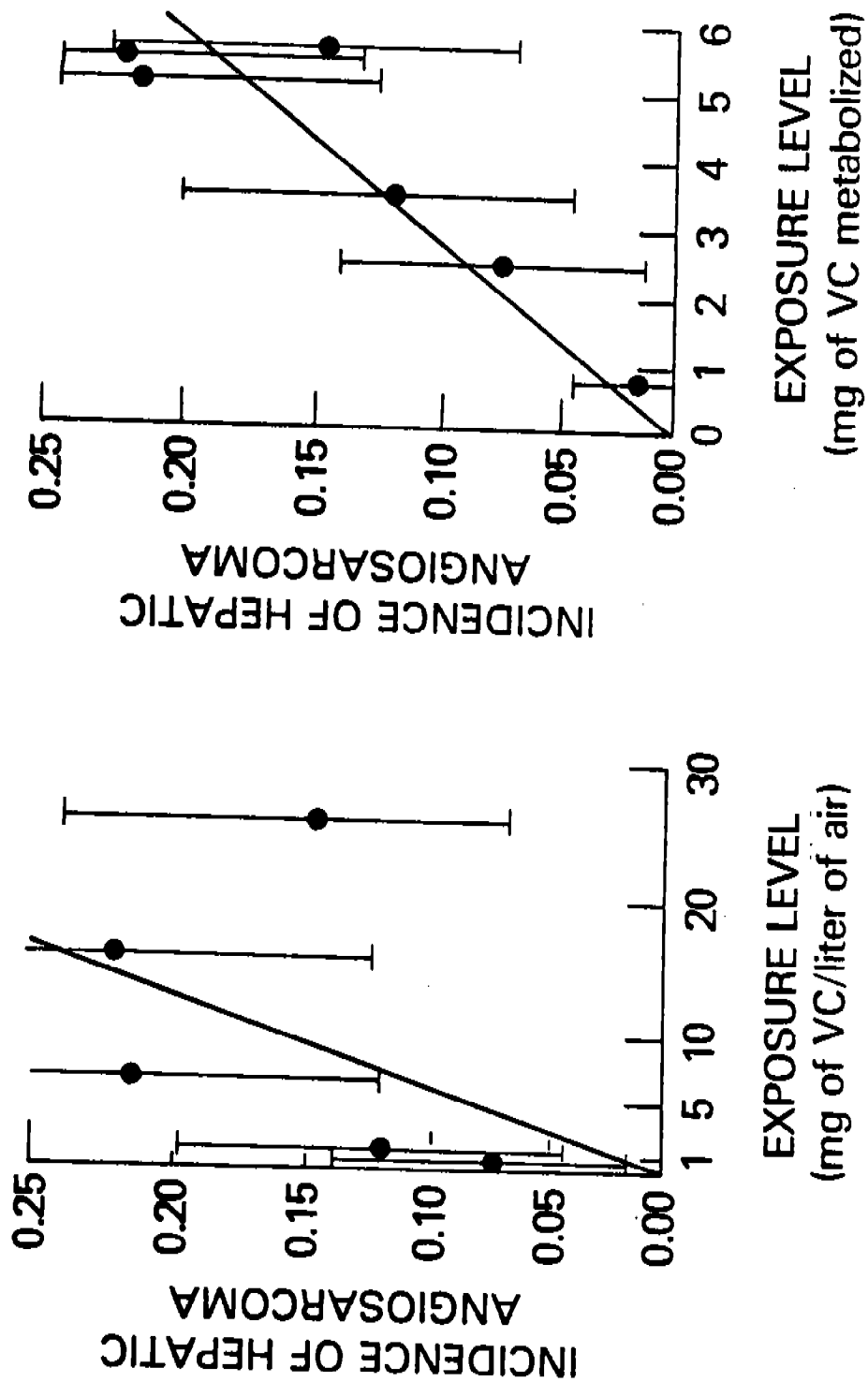


FIGURE 6: Vinyl chloride induced hepatic angiosarcomas in rats
[data from Gehring et al. (29)]

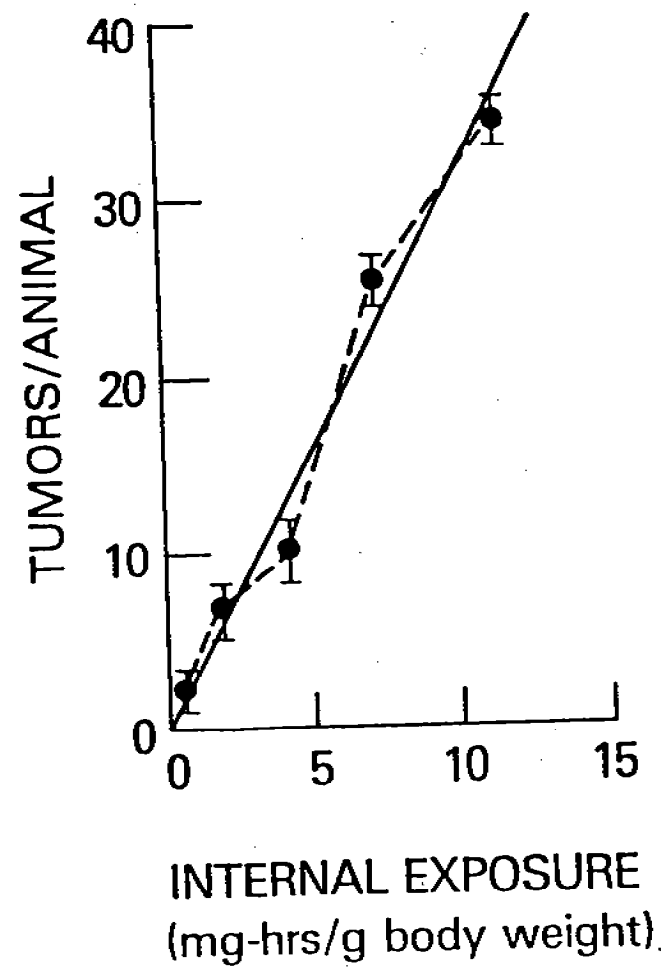
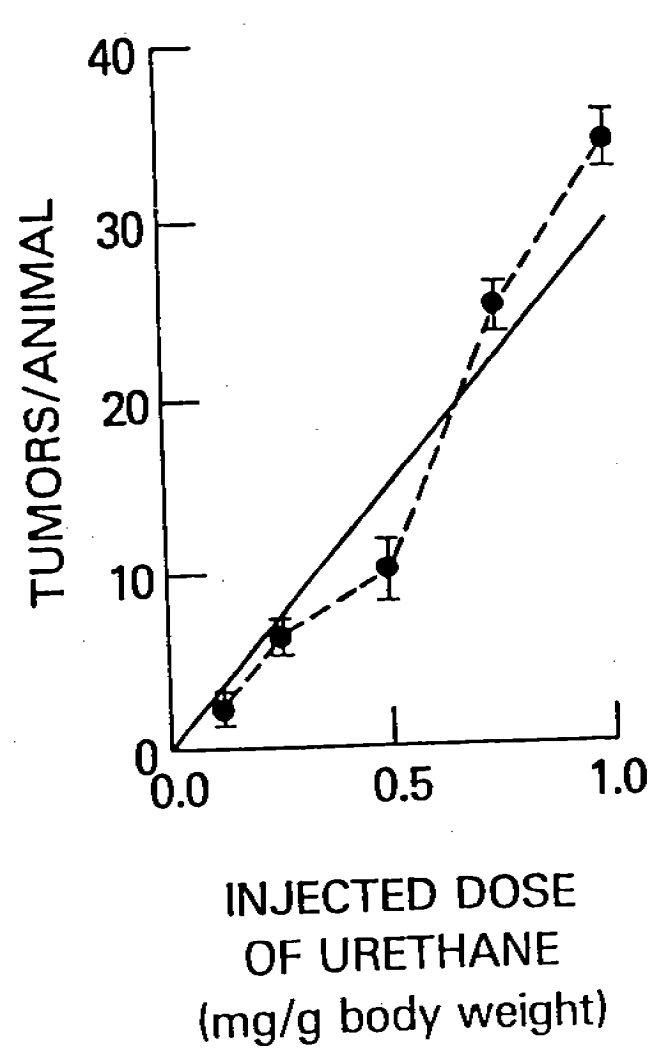


FIGURE 7: Urethane induced lung adenomas in mice
[data from White (30)]

**Twenty-ninth Conference on the Design of Experiments in Army Research,
Development, and Testing (19-21 October 1983)**

Attendees

Mark S. Adams
 Harold Ascher
 William E. Baker
 Richard M. Bates
 Barney Bissinger
 Barry A. Bodt
 Dimitrios T. Boumpos
 Dean Bross
 John Brundage
 J. Robert Burge
 Gregory Campbell
 Linda L. Crawford
 David F. Cruess
 H.A. David
 Douglas J. DePriest
 Mary Dufour
 Lane Felker
 Walter D. Foster
 Greg Gibson
 Alfred D. Godfrey
 Ian Gordon
 Frank E. Grubbs (Retired)
 Jock O. Grynovicki
 Bernard Harris
 Ellen Hertz
 Herbert H. Holman
 John Holter
 Michio Horigome
 Jacqueline A. Horton
 Russell E. Hudson
 Charles Hunter

 Alan Johnsrud
 Martin W. Joseph
 A.A. Kahn
 Eve Kaplan
 Leon Katz
 Thomas Kitchell
 Jim Knaub
 Richard Kyle
 Les Lancaster
 Robert L. Launer
 Charles R. Leake
 James A. Lechner
 Gary C. Love
 John Lyons

Organization

U. S. Army Concepts Analysis Agency
 Naval Research Laboratory
 Armament Research & Development Ctr (BRL)
 DoD
 Dept. of Mathematical Sciences - Penn State U.
 Ballistic Research Laboratory
 NIH
 U of MD
 WRAIR, WRAMC, Washington, DC
 Walter Reed Army Medical Center
 NIH, Bethesda, MD
 Armament Research & Development Ctr (BRL)
 Div. of Biostatistics, USUHS
 (Iowa State Univ) Statistical Laboratory
 Office of Naval Research
 DBE/NIAAA
 HUD, Washington, DC
 Frederick, MD
 US Army Systems Analysis
 Washington, D.C.
 NIH
 US Army Materiel Systems Analysis Activity
 Armament Research & Development Ctr (BRL)
 Mathematics Research Ctr.-Univ. of Wisconsin
 USA OTEA
 DoD, Ft. Mende
 Aberdeen Proving Ground
 Dept. of Operations Research/GWU
 Dept. of PrevMed/Bio, USUHS
 Social Security Administration
 Operational Research & Analysis Establishment -
 Dept of National Defense, Canada
 Concepts Analysis Agency
 US Army Support, St. Louis, MO
 U.S. ACAA
 WRAIR
 US Army Personnel Cntr, Alexandria, VA
 Silver Spring, MD
 US Army Logistics Center
 Dept. of Anesthesiology, USUHS
 US Nuclear Regulatory Commission
 Army Research Office
 Springfield, VA
 National Bureau of Standards
 Ft Ord, CA, US Army Combat Development
 Applied Physics Lab, Johns Hopkins

Donald MacCorguodale
Dale Madden
John Mandel
John G. Mardo
Richard T. Maruyama
Michael McGrath
William McIntosh
Jules Merkler
Paul Michael
Alexande Mickiewicz
David Miller
Ed Milton
J. Richard Moore
Craig R. Morrisette
David R. Musser
Alice I Nichols
Sam Padula

Emanuel Parzen
Robert Paule
Karen Pettigrew
Ronald P. Reale
Edward W. Ross
Carl T. Russell
William Sacco
Jerome Sacks
James J. Schlesselman
Paul J. School
Rudolph Schwartz
Richard H. Scissors
Seymour M. Selig
Ellen Shannahan
Helen Sing
N.D. Singpurwalla
Dennis Smith
J. Smith
Robert Smythe
Cliff Speigelman
Raymond Spring
Andrew Soms
Refik Soyer
Perry Stewart
Donald M. Swingle
Douglas B. Tang
Malcolm S. Taylor
Joseph M. Tessmer
Jerry Thomas
Linda Tompkins
Marcus A. Weinberger

Daniel Willard
Lang Whitters
Michael Woodrooffe
Jim Y. Yen

Aberdeen Proving Ground
FDA
National Bureau of Standards
USA ARDC
DCS Army Test & Evaluation Command
DoD, The Pentagon, Off. of Sec of Defense
US Army Test & Evaluation Command
Aberdeen Proving Ground
DoD, Ft. Meade
Aberdeen Proving Ground
US Army (OTEA)
US Army (OTEA)
Ballistic Research Lab
Walter Reed Army Medical Center
USA DAR COM HQ, Alexandria, VA
USUHS
Aberdeen Proving Ground, Director, HMSAA
OEMD

Texas A & M University
National Institutes of Health
National Institutes of Health
USA Concepts Analysis Agency
Aero Mechanical Engineering Lab
OTEA
Bellaire, MD
National Science Foundation
Div. of Biostatistics, USUHS
Ft Belvoir, VA
Social Security Administration
Silver Spring, MD
Office of Naval Research
US Army Infantry School
Walter Reed Army Medical Center
Dept. of Operations Research

Armament Research & Development Ctr (BRL)
AFOSR/NM
National Bureau of Standards
US Army Natick Research & Development
Math Research Ctr.
Arlington, VA
US Army, Ft Lee, VA
Las Cruces, New Mexico
Dept. of Biostatistics, WRAIR
Ballistic Research Lab
USAF/SAGF
Ballistic Research Lab
USA OTEA
Operational Research & Analysis Establishment,
Canada
Office, Deputy Under Secretary of the Army
USA OTEA
U of Michigan and Rutgers University
Ft Ord, CA (SSSC/SSL/CDEC)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO Report 84-2	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PROCEEDINGS OF THE TWENTY-NINTH CONFERENCE ON THE DESIGN OF EXPERIMENTS IN ARMY RESEARCH, DEVELOPMENT AND TESTING		5. TYPE OF REPORT & PERIOD COVERED Interim Technical Report
7. AUTHOR(s)		6. PERFORMING ORG. REPORT NUMBER
9. PERFORMING ORGANIZATION NAME AND ADDRESS		8. CONTRACT OR GRANT NUMBER(s)
11. CONTROLLING OFFICE NAME AND ADDRESS Army Mathematics Steering Committee on Behalf of the Chief of Research, Development and Acquisition		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) US Army Research Office P. O. Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE June, 1984
		13. NUMBER OF PAGES 356
		15. SECURITY CLASS. (of this report) UNCLASSIFIED
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited. The findings in this report are to be construed as Official Department of the Army position, unless so designated by other authorized documents.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES This is a technical report from the Twenty-ninth Conference on the Design of Experiments in Army Research, Development and Testing. It contains most of the papers presented at that meeting. These articles treat various Army statistical and design problems.		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) injury severity scoring Boyes sampling theory sequential Bernoulli selection quantal response probability distribution types normal and t tail probabilities complex demodulation cycles of suicide optical data operational testing reliability random numbers bootstrap methods missing data sequential testing fire support complex computer model carcinogenesis studies		

