ENHANCING HUMAN PERFORMANCE

Issues, Theories, and Techniques

Daniel Druckman and John A. Swets, Editors

Committee on Techniques for the Enhancement of Human Performance
Commission on Behavioral and Social Sciences and Education
National Research Council

NATIONAL ACADEMY PRESS
Washington, D.C. 1988

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

NATIONAL ACADEMY PRESS • 2101 Constitution Avenue, NW • Washington, DC 20418

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This report has been reviewed by a group other than the authors according to procedures approved by a Report Review Committee consisting of members of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine.

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Samuel O. Thier is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

Library of Congress Cataloging-in-Publication Data

Enhancing human performance : issues, theories, and techniques / Daniel Druckman and John A. Swets, editors.
p. cm.
"Committee on Techniques for the Enhancement of Human Performance, Commission on Behavioral and Social Sciences and Education, National Research Council."
Bibliography: p.
Includes index.
ISBN 0-309-03792-1. ISBN 0-309-03787-5 (soft)
1. Self-realization—Congresses. 2. Performance—Psychological aspects—Congresses. I. Druckman, Daniel, 1939– . II. Swets, John Arthur, 1928– . III. National Research Council (U.S.). Committee on Techniques for the Enhancement of Human Performance.
BF637.S4E56 1987
158—dc19 87-31233
CIP

Printed in the United States of America

COMMITTEE ON TECHNIQUES FOR THE ENHANCEMENT OF HUMAN PERFORMANCE

JOHN A. SWETS, *Chair,* Bolt Beranek and Newman Inc., Cambridge, Mass.
ROBERT A. BJORK, Department of Psychology, University of California, Los Angeles
THOMAS D. COOK, Department of Psychology, Northwestern University
GERALD C. DAVISON, Department of Psychology, University of Southern California
LLOYD G. HUMPHREYS, Department of Psychology, University of Illinois
RAY HYMAN, Department of Psychology, University of Oregon
DANIEL M. LANDERS, Department of Physical Education, Arizona State University
SANDRA A. MOBLEY, Director of Training and Development, The Wyatt Company, Washington, D.C.
LYMAN W. PORTER, Graduate School of Management, University of California, Irvine
MICHAEL I. POSNER, Department of Neurology, Washington University
WALTER SCHNEIDER, Department of Psychology, University of Pittsburgh
JEROME E. SINGER, Department of Medical Psychology, Uniformed Services University of Health Sciences, Bethesda, Md.
SALLY P. SPRINGER, Department of Psychology, State University of New York, Stony Brook
RICHARD F. THOMPSON, Department of Psychology, Stanford University

DANIEL DRUCKMAN, *Study Director*
JULIE A. KRAMAN, *Administrative Secretary*

Contents

Preface

The Army Research Institute in 1984 asked the National Academy of Sciences to form a committee to examine the potential value of certain techniques that had been proposed to enhance human performance. As a class, these techniques were viewed as extraordinary, in that they were developed outside the mainstream of the human sciences and were presented with strong claims for high effectiveness. The committee was also to recommend general policy and criteria for future evaluation of enhancement techniques by the Army.

The Committee on Techniques for the Enhancement of Human Performance first met in June 1985. The 14 members of the committee were appointed for their expertise in areas related to the techniques examined. The disciplines they represent include experimental, physiological, clinical, social, and industrial psychology and cognitive neuroscience; one member is a training program director from the private sector. During the next two years, the committee gathered six times, met in toto or in part on several occasions with various representatives of the Army, conducted interviews and site visits and sent subcommittees on several others, and commissioned 10 analytical and survey papers. The committee also examined a variety of materials, including state-of-the-art reviews of relevant literature, reports commissioned by the Army Research Institute, and unpublished documents provided by institutes, practitioners, and researchers. The report that follows describes the committee's activities, findings, and conclusions. Though cast largely in terms of the sponsor's setting, this report is relevant to other settings, for example, industry. The next few paragraphs present some background.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

That the United States Army should be concerned to enhance the performance of its personnel is self-evident. We know that young volunteers must become not only soldiers who do well in battle but also technicians who skillfully operate and maintain complex equipment in peace and war. We are aware, moreover, that personal skills are not enough: individuals are heavily dependent on each other within small groups, and groups of various sizes must work very effectively together to permit survival and ensure success. And, of course, all must be ready to give peak performances in situations of great hardship, uncertainty, and stress. In the face of these staggering requirements, one must realize that turnover of personnel is high and that the training time available—to impart the necessary cognitive, physical, and social skills—is brief.

So it comes as no surprise that the Army is on the lookout for techniques that can help enhance human performance. The Army Research Institute is charged with seeking out and developing such techniques: it does so by employing researchers in the human sciences and by supporting appropriate research in universities and other public and private organizations. It focuses largely on promising new techniques as they appear in the mainstream of behavioral, physiological, and social research. However, given the pressures and given a view of mainstream research as slow, narrow, and insufficiently targeted, it also comes as no surprise that some influential officers and certain segments of the Army want to cast a broader net to snare promising enhancement techniques. To do this, they look beyond traditional research organizations and practices to what are viewed as extraordinary techniques. These techniques are thought possibly to provide such unusual benefits as accelerated learning, learning during sleep, superior performance through altered mental states, better management of behavior under stress, more effective ways of influencing other people, and so on. There is also an initiative within the Army to consider techniques based on paranormal phenomena, for example, extrasensory perception to view remote sites and psychokinesis to influence the operation of distant machines.

Along with these urgings to examine, to try, or to implement extraordinary techniques come difficult new problems for those in the Army responsible for evaluation, as well as for those in the Army responsible for personnel and training practices. One issue is that proponents of such techniques are usually not content with traditional evaluation procedures or scientific standards of evidence, often giving more weight to personal experience and testimony. Furthermore, a typical technique of this kind does not arise from the usual research traditions of experiments published in refereed journals and peer review of cumulated evidence, but rather appears full-blown as a package promoted by a commercial vendor. What does the Army Training and Doctrine Command or the base commander

do when the need is great, the package is ready, the claims are for miracles, some senior officers are vocally supportive, and the evaluation criteria are fluid? What do Army intelligence agencies do when the same conditions apply and other nations are said to be active in investigating paranormal effects?

The committee decided to assess a representative set of the techniques in question and resolved to address the surrounding issues in an open-minded and thorough way. We therefore divided ourselves into a number of subcommittees organized according to the behavioral processes addressed by the several techniques: accelerated learning, sleep learning, guided imagery, split-brain effects, stress management, biofeedback, influence strategies, group cohesion, and parapsychology. In addition, a subcommittee on evaluation issues was formed to examine practices and standards relevant to all the techniques. Each chapter of the report was prepared by the appropriate subcommittee, but interactions were frequent and so the report represents a collaborative effort of all the members.

Chapter 1 provides a context for the committee's task and the Army's interest in enhancing performance, characterizes some particular techniques, and introduces some general issues in evaluating them. Chapter 2 presents the committee's findings about the techniques examined and conclusions about appropriate evaluation procedures. Chapter 3 treats the relevant evaluation issues more systematically and presents the committee's philosophy of evaluation as it pertains to the matter at hand. Chapters 4 through 8 deal with particular techniques but are organized in terms of more general psychological processes. Chapter 9 considers parapsychological techniques.

The report concludes with six appendixes. Appendix A briefly summarizes the key elements of each enhancement technique. Appendix B lists the ten papers commissioned by the committee and their authors. Appendix C lists the members and activities of the subcommittees and also the activities of the committee as a whole. Appendix D lists key terms used in the research on particular techniques. Appendix E discusses the application of scientific research by the military. Appendix F contains biographical sketches of the committee members.

As committee chair, I am now in the pleasant position of recounting the several contributors to the total committee process, a process that went remarkably well. Definition and guidance for the committee's task came primarily from Edgar M. Johnson, director of the Army Research Institute. Administrative and technical liaison was ably provided by project monitor George Lawrence, who worked closely with the committee in its various activities. They were supported well by several senior Army officers, including Colonel William Darryl Henderson, Commander of the Army Research Institute; Major General John Crosby,

Assistant Deputy Chief of Staff for Personnel; and General Maxwell R. Thurman, Vice Chief of Staff. The committee met with members of a resource advisory group that included Lieutenant General Robert M. Elton, chair, Deputy Chief of Staff for Personnel; Lieutenant General Sidney T. Weinstein, Assistant Chief of Staff for Intelligence; Dr. Louis M. Cameron, Director of Army Research and Technology; Major General Maurice O. Edmunds, Commander of the Soldier Support Center; and Major General Philip K. Russell, Commander of the Medical Research and Development Command. Among the Army staff who were very helpful to the committee are Colonel John Alexander and Mr. Robert Kraus; the names of many others appear in Appendix C.

The committee's two consultants contributed special expertise: Paul Horowitz (of Bolt Beranek and Newman Inc.) joined the site visits of the subcommittee on parapsychology and advised on physical aspects of experiments in that area; James Schroeder (of Southwest Research Institute) attended the committee's meeting at Fort Benning, Georgia, and advised on the application of scientific research by the military (see Appendix E). The committee also received special expertise by commissioning papers. These papers and their authors are listed in Appendix B.

At the National Research Council, David Goslin, executive director of the Commission on Behavioral and Social Sciences and Education, once again provided wise counsel and support. Ira Hirsh, commission chair, and William Estes, also representing the commission, gave valuable advice and encouragement. Thomas Landauer, a member of the NRC's Committee on Human Factors, provided liaison in the areas of our committees' mutual interests. The reviewers of this report gave us a good measure of reinforcement along with helpful critiques. Eugenia Grohman, associate director for reports, lent experience and wisdom to this report. Special gratitude is extended to Christine McShane, the commission's editor: her skillful editing of the entire manuscript contributed substantially to its readability, and the coherence of the volume owes much to her suggestions for organizing the material. Julie Kraman, as administrative secretary to the committee, earned its considerable appreciation for setting up efficient meetings and for handling all manner of tasks graciously and smoothly.

Daniel Druckman, study director of the project, receives the committee's great appreciation for his intellectual contributions across the broad range of topics considered as well as for his logistic support. Working closely with the authors of chapters and commissioned papers, he provided an integration of the several contributions as well as much of the introductory and interstitial material. He also served on two subcommittees in areas of his expertise.

The ultimate debt of anyone who finds this report useful, and my large

personal debt, is to the members of the committee. As individuals, their capabilities are broad and deep. As a group, they gave generously and productively of their time, were always engaged, responded to every challenge, and, especially, showed an exceptional talent for reaching consensus in a collegial, advised, and efficient way.

JOHN A. SWETS, *Chair*
Committee on Techniques for the
Enhancement of Human Performance

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

PART I

Overview

PART I CONSISTS OF THREE CHAPTERS. Chapter 1 sets the stage for the report. It describes the committee's task, provides background on the Army's interest in enhancement techniques, characterizes specific techniques examined by the committee, and identifies the main issues in evaluating the relation between techniques and human performance. Chapter 2 presents the committee's findings and conclusions. We draw general conclusions about the process of consideration given to any technique and state specific findings and conclusions for each of the areas of human performance examined.

Chapter 3 presents the committee's philosophy of evaluation as it pertains to enhancement techniques. Some of the issues involved concern the conduct of basic research; others concern the conduct of field tests. With respect to basic research, issues include the plausibility of inferences about novel concepts, causation, alternative explanations of causal relations, and the generalizability of causal relations. With respect to field tests, a number of questions are of interest: Does the enhancement program meet genuine Army needs? Is the resulting program implementable, given program design and resources? Do unintended side effects limit utility? Is the program more cost-effective than its alternatives? These questions underscore the reality that evaluation research is largely a pragmatic activity influenced by the organizational context in which it occurs.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

1

Introduction

THE COMMITTEE'S TASK

At the request of the U.S. Army Research Institute, the National Research Council formed a committee to assess the field of techniques that are claimed to enhance human performance. The Institute asked the Council to evaluate the claims made by proponents of selected existing techniques and to address two general additional questions: (1) What are the appropriate criteria for evaluating claims for such techniques in the future? (2) What research is needed to advance our understanding of performance enhancement in areas related to the proposed techniques? The objectives of the committee's study are to provide an authoritative assessment of these questions for policymakers in research and development who are consumers of the techniques, as well as to consider their possible applications to Army training.

Many of the techniques under consideration grew out of the human potential movement of the 1960s, including guided imagery, meditation, biofeedback, neurolinguistic programming, sleep learning, accelerated learning, split-brain learning, and various techniques to reduce stress and increase concentration. Many of these techniques have gained popularity over the past two decades, promoted by persons eager to provide answers to problems of human performance or to prosper from them. While often using the language of science to justify their approach, these promoters are for the most part not trained professionals in the social and behavioral sciences. Nonetheless, they do appeal to basic needs for human performance, and the Army, like many other institutions, is attracted to the prospect of cost-effective procedures that can improve performance.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

These institutions must evaluate the effects of such procedures, however. Issues include the appropriateness of a quick-fix approach, the distinction between the impact of an experience and actual change, and the plausibility of evidence indicating that something is happening even if the effects are not reproducible or the benefits uncertain.

A more conservative atmosphere in the 1980s is reflected in the way techniques are advanced. Motivation in the 1980s may be primarily entrepreneurial, not ideological, as it was in the 1960s. Advocates focus on relating the techniques to specific tasks, such as marksmanship, foreign language acquisition, fine motor skills, sleep inducement, and even combat effectiveness. Some techniques are in fact rooted in a scientific literature. For these reasons the various techniques have attracted the interest of institutions that have rejected, and would probably continue to reject, countercultural trends in society. Indeed, much attention has been given to these techniques by industrial, government, and military policymakers, as well as by the general public. For this reason especially, it is important to address the issues surrounding the claims made for effectiveness.

Elaborate training programs have grown, nourished by their developers' enthusiasm and salesmanship in a social context receptive to quick cures. For many of these programs, success in the marketplace is used to justify the approaches. For others, more esoteric concepts, including the role of neurotransmitters, the physics of neuromuscular programming, brain wave patterns, hemispheric laterality, high-access memory storage, preferred sensory modalities, and low-gain innervation of muscles, are used to attempt to provide scientific justification for the claims. The chapters that follow evaluate the evidence and theories used to support the claims of several popular techniques. Before turning to these evaluations, however, we provide some background on the Army's interest in these techniques, as well as a discussion of issues surrounding enhanced performance and issues in evaluating the relation between techniques and performance.

THE ARMY'S NEEDS

The Army motto, "Be all that you can be," symbolizes the current ethos of the institution, an army of excellence. Emphasis is placed on attaining certain ideals, such as fearlessness, cunning, courage, one-shot effectiveness, fatigue reversal, and nighttime fighting capabilities. These ideals are assumed to be realizable through training, even if the most effective techniques have not as yet been identified. The culture of improvement is further reinforced by the dilemma created by an all-volunteer Army and the demands of complex new computer technologies. Many civilians enter military service with only the required minimum of

formal education; most of these volunteers enlist in the Army. For this reason, the Army's emphasis on skill training is well founded.

The importance of the human element in combat is recognized in the Army Science Board's 1983 report "Emerging Concepts in Human Technology," which phrases the issue in terms of high yield at relatively low investment. Human capital is considered to be the best potential source for growth in Army effectiveness, both in terms of return on investment and as a moral imperative "if we are to commit our soldiers to fight outnumbered and win." The technologies singled out in the report are those that can improve creativity and innovation, learning and training, motivation and cohesion, leadership and management, individual, crew, and unit fitness, soldier-machine interface, and the general productivity of the Army's human resources.

The Board's report largely bypasses issues of systematic evaluation of enhancement techniques within the Army context, while addressing mechanisms for integrating them with Army activities. Little concern is shown for adducing relevant criteria to determine whether implementation is feasible. The Army's ambitious goals, combined with a reluctance to deal with the complexities surrounding issues of human performance, make this institution potentially susceptible to a variety of claims made by technique developers. It would therefore seem prudent to devise criteria for evaluating those claims.

A SELLER'S MARKET

Techniques for enhancement of human performance have received much attention in the popular press. They have been actively promoted by entrepreneurs who sense a profitable market in self-improvement. The American Society for Training and Development "estimates that companies are spending an astounding $30 billion a year on formal courses and training programs for workers. And that's only the tip of the iceberg" (*Wall Street Journal,* August 5, 1986). They are also taken seriously by the U.S. military, who are at times accused of losing the "mind race" to the Soviets (see, for example, Anderson and Van Atta, *Washington Post,* July 17, 1985). The Army has shown particular interest in techniques that help people acquire, maintain, or improve such skills as classroom learning, communication and influence, creativity, and accuracy in the execution of tasks requiring motor skills. Those that are cost-effective and produce relatively rapid results are likely to receive the most attention, along with research breakthroughs that could be a basis for new training programs. What are these techniques? What claims are being made for them? Is there evidence that substantiates these claims?

Examples of techniques include biofeedback (information about internal

processes), Suggestive Accelerative Learning and Teaching Techniques (a package of methods geared primarily toward classroom learning), hemispheric synchronization (a machine-aided process based on assumptions about right brain–left brain activities), neurolinguistic programming (procedures for influencing another person), and Concentrix (a procedure used to improve concentration on specific targets). Also of interest to the Army are such processes as group cohesion and stress reduction, as well as the claims for sleep learning, peak performance, and parapsychology. Together, these techniques and processes cover the major types of skills—motor, cognitive, and social. Several of them are described here briefly, along with illustrative claims found in brochures and course material.

Suggestive Accelerative Learning and Teaching Techniques (SALTT) is an approach to training that employs a combination of physical relaxation, mental concentration, guided imagery, suggestive principles, and baroque music with the intent of improving classroom performance. Some applications have included language training, typing instruction, and high school science courses. Attempts have been made to evaluate the applications, and many of these evaluations are published in the *Journal of the Society for Accelerative Learning and Teaching* (Psychology Department, Iowa State University). The following is a sampling of claims made in brochures and convention announcements: "A proven method which has broad potential application in U.S. Army training"; "It will significantly reduce training time, improve memory of material learned and introduce behavioral changes that positively affect soldier performance—self-esteem, self-confidence, and mental discipline"; and "Most students will prove to themselves that they have learned a far greater amount of material per unit of time with a greater amount of pleasure than they have ever previously done."

Neurolinguistic programming (NLP) refers to a set of procedures developed to influence and change the behaviors and beliefs of a target person. Its goals are mostly therapeutic, but its proponents also advocate the use of the techniques in advertising, management, education, and interpersonal activities. A small research literature, published primarily in the *Journal of Counseling Psychology*, has developed. Practitioners can be trained and certified at various institutes, and the National Association for Neurolinguistic Programming distributes a newsletter to its membership, currently about 500 persons. Illustrative claims and testimonials found in advertising materials include: "[NLP] has evolved a unique technology which encompasses a set of specific techniques enabling you to produce well-defined results" and "NLP . . . is clear, easy to learn, and brilliant." A typical slogan is that found in a brochure from the Potomac Institutes, Silver Spring, Maryland: "The difference

that makes the difference, for education, management, psychotherapy, psychiatry, business, law, health care, and the arts."

Hemi-Sync™, which is short for hemispheric synchronization, is a technique that consists of presenting two tones slightly differing in frequency to separate ears with stereo headphones to produce binaural beats. The long-known result is a tone that waxes and wanes at a frequency equal to the difference between the original tones. Pioneered as an enhancement technique by Robert Monroe of the Monroe Institute of Applied Science in Faber, Virginia, the technique is based on the assumption of a frequency following response (FFR) in the human brain. The FFR refers to a correspondence between sound signals heard by the ear and electrical signals recorded by an electroencephalograph (EEG). It is claimed that, by altering sound patterns, it is possible to alter states of awareness. Stated applications are in the areas of language learning, stress management, reading skills, and creativity and problem solving. Claims of effectiveness stated in the Monroe Institute's brochure are wide-ranging, covering education (e.g., "77.8 percent of a class reported improvement in mental-motor skills"), health (early recuperation, lower blood pressure), psychotherapy (stress reduction, working with terminally ill patients, teaching autistic children), and sleep restorative training (e.g., "forty of forty-five insomniacs reported that one-month use of Hemi-Sync™ tapes was at least as effective as medication, without the drug side effects").

SyberVision® is a scripted videotape that presents an expert (e.g., a world-class athlete) repeatedly performing fundamental skills of his or her activity (e.g., golf) without verbal instructions. It is based loosely on principles of vicarious learning, guided imagery, and mental rehearsal. Developed and marketed by SyberVision Systems Inc., San Leandro, California, the package includes a cassette and instruction manual with an appendix on the "simple physics of neuro-muscular programming." The appendix presents a scientific rationale for the technique, for example, "the more you see and hear pure movement, the deeper it becomes imprinted in your nervous system . . . and the more likely you are to perform it as a conditioned reflex," and "The decomposition of what is seen and sensorily experienced into an electromagnetic wave form is accomplished by a complex mathematical operation (Fourier Transform) by the brain" (*Instruction Manual on Golf with Patty Sheehan*). Support for enhanced performance is, however, based on testimonials rather than experiments, for example, Killy on skiing, a Stanford tennis coach on tennis, Professional Golf Association members on golf, Peters (*In Search of Excellence*) on achievement, Salk on leadership, and a variety of corporate executives and educators on self-improvement. Claims range from sweeping statements (e.g., "We owe these two men a large debt of

gratitude") to rather precise statements (e.g., "In 47 days I have lost 25 pounds [191 to 166], yet I look like I lost 40") (in the United Airlines magazine, *Discoveries*). This technique involves a significant marketing effort that builds on users' willingness to be quoted and the use of acknowledged academic experts (e.g., Stanford neuropsychologist Karl Pribram), whose role in the program is advertised as being central.

Stress management techniques are procedures designed to alleviate anxiety or tension. Catering to an age of anxiety, self-help books, groups, and clinics on managing stress proliferate. A good example of the approach is the recent book by Charlesworth and Nathan (1982), which emphasizes fitness, nutrition, managing time, general life-styles and life-cycles, as well as strategies such as progressive relaxation, autogenic training, and image rehearsal. Appendixes provide the reader with home practice charts, a guide to self-help groups, and suggested books and recordings. The groups offer their members information, emotional support, and a sense of belonging. Often stress management procedures are combined with a number of other techniques into a single package. The promoters often emphasize the total package rather than particular techniques; the packages usually combine several processes that, when acting together, are thought to produce significant effects.

The Army's needs for techniques that can improve performance make it subject to the sorts of claims illustrated above. While they and other consumers can avoid the more obvious pitfalls, the proliferation of choices and products and the lack of scientific evidence allow marketplace criteria to become the bases for decisions. But there are exceptions. Some techniques have received the attention of the scientific community, and evidence is available to be used as criteria in such areas as biofeedback, guided imagery, sleep learning, cohesion, and even for some aspects of psychic phenomena and neurolinguistic programming.

The literature has alerted us, for example, to the distinction between the effects of biofeedback on fine motor skills and on stress, to the different effects of mental and physical rehearsal, to placebo and Hawthorne effects in stress research, to the priming and repetition effects of material presented during sleep, to some dysfunctions of group cohesion, to the difficulties of replicating experiments on extrasensory perception, and to the implausibility of specialized sensory modalities as postulated by NLP (see Appendix D for key terms). These findings make evident a complex relation between technique and performance.

IMPROVED PERFORMANCE: COMPLEX ISSUES, SIMPLE SOLUTIONS

The research literature in such traditional areas of experimental psychology as learning, perception, sensation, and motivation suggests

complex relations between interventions and improved performance. Many technique promoters appear to pay little attention to this literature, preferring an alternative route to invention: rather than derive a procedure from appropriate scientific literature, they create techniques from personal experiences, sudden insights, or informal observation of "what works." Science may enter the process after the technique is developed and used, for example, to legitimize its use or to endorse methods for evaluation. Research follows rather than precedes the invention. This sequence increases the likelihood that important considerations will be missed. We highlight some of these considerations in this section.

The lack of easy avenues to improved performance may well be due to the complexity of the behavior in question. One definition of skills emphasizes the importance of the coordination of behavior: "A skilled response . . . means one in which receptor-effector-feedback processes are highly organized, both spatially and temporally. The central problem for the study of skill learning is how such organizations or patterning comes about" (Fitts, 1964:244). This definition implies that skill learning involves an orchestration of diverse processes, making the topic an interesting one to various subfields of psychology. It also makes evident a number of unresolved issues, including whether different skills are learned and retained in different ways. The research findings obtained in this literature contribute to our understanding of the necessary, if not sufficient, conditions for improved performance.

Research on skill acquisition addresses such basic questions as What are the stages of learning? and What is learned? Distinctions made between short-term and long-term memory storage and between schemas and details have contributed to our understanding of basic processes (see Welford, 1976). Other questions have more direct consequences for application: for example, what contributes to the acquisition and maintenance of skills? How can the adverse effects of stress, fatigue, and monotony be avoided? These questions are the basis for programs of research that can be divided into several parts, each defined in terms of empirical issues (Irion, 1969; see also the other chapters in Bilodeau and Bilodeau, 1969). Some examples of empirical issues are practice effects (differences due to distributed versus massed practice, long versus short rest periods, short versus long sessions), the whole-part problem (differences due to learning a task as a whole versus learning it by its constituent elements), feedback (differences due to delays in receiving knowledge of results and to type of information during the delay period), retention (differences due to whether the the task is motor or verbal), and transfer of training.

These and related considerations suggest that skill learning is an incremental process likely to differ from one type of skill to another. Whether intending to enhance motor, verbal, problem-solving, or social

performances, technique designers can ill afford to ignore these lessons from the experimental literature on skill acquisition and maintenance. It is also the case, however, that the agenda of unexplored issues is much larger than the accomplishments to date, and this is recognized particularly in the rapidly growing field of cognitive psychology, in which the "information-processing revolution" is just beginning.

Practical applications are, however, not automatic. Many excellent applications do not spring from basic science; some are the result of craft and experience. More important perhaps are the indirect contributions made in both directions—from basic to applied and vice versa. A systematic approach taken in both domains serves to vitalize each, as when applied investigations reveal new phenomena that need explanation or when a new package incorporates basic principles discovered originally in the laboratory. Such an approach is likely to facilitate the design of appropriate techniques for skill acquisition. At issue is whether a particular technique can produce and sustain desired changes.

One conclusion from the research accumulated to date is that effective interventions are those that are continuous and self-regulating and take account of both context and person (see, for example, Lerner, 1984). Particularly relevant is the difference between short-term and long-term changes. Effects obtained by many techniques for performance enhancement may be short-term in their effects. This distinction is made by Back (1973, 1987) in his evaluation of the sensitivity training movement. The changes observed by sensitivity trainers and documented by evaluators may well reflect the impact of the experience per se. Such situation effects are unlikely to be sustained in different environments, an observation supported by the literatures in both developmental and social psychology (Druckman, 1971; Frederiksen, 1972). These literatures caution against hasty generalizations from observed, situation-specific effects; they also explain why long-term effects may be difficult to produce with brief exposures to "treatments." Like the sensitivity trainers of the 1960s and 1970s, many of the promoters (and consumers) of the 1980s pay little attention to issues of causality and intrinsic motivation, preferring instead to dwell on single dimensions of treatments or to offer a mixed package constructed in arbitrary ways and producing diffuse effects that reflect the experience.

The issue of expected benefits from techniques provides a bridge between research and application. Research can be designed to evaluate techniques, as well as to discover possible unintended side effects. Indeed, a research literature has developed in some of the areas examined in this book, namely biofeedback, stress, and guided imagery. For many other techniques, however, a relevant body of research does not exist; this lack applies to some of the techniques examined by the committee,

as well as to those yet to appear on the market. It is these techniques that present a problem for us as evaluators. Evaluation without data is difficult, but not impossible. Our approach is to place the techniques into broader categories corresponding to the key processes being influenced, for example, learning, motor skills, and influence. By so doing, the claims can be evaluated within the frameworks of existing theories and methodologies. They can also be judged against results obtained in related areas. This approach serves as the organizing theme for the chapters that follow.

EVALUATING THE TECHNIQUES

Evaluations properly hinge on answers to a standard set of questions proposed in a paper entitled "Evaluating Human Technologies: What Questions Should We Ask?" by Hegge, Tyner, and Genser (1983) at the Walter Reed Army Institute for Research:

- What changes will the technique produce?
- What evidence supports the claims for the technique?
- What theories stand behind the technique?
- Who will be able to use the technique?
- What are the implications of the technique for Army operations?
- How does the technique fit with Army philosophy?
- What are the cost-benefit factors?

These questions served as guidelines for the committee's evaluations. Appendix A is a summary description of each technique, organized along the lines of the Hegge, Tyner, and Genser questions, covering theory, research, and application. For many of the categories, however, the desired information is either too limited to be useful or simply not available; in such cases we have considered other strategies for evaluation.

The committee faced a number of difficulties in evaluation that stem from recurrent problems posed by the technologies. One is the tendency for some promoters (and consumers) to rely primarily on testimonials or anecdotal evidence as a basis for application. Another is a general lack of strong research designs to provide evidence of effects. These problems are considered also in the context of specific techniques discussed in the chapters of Parts II and III.

Practitioners of techniques often emphasize the value of personal or clinical experience and marketplace popularity as bases for judging the techniques. They are generally less inclined to seek research evidence or to support research evaluation programs. These attitudes may be related to the fact that few practitioners are trained as researchers. For some it is sufficient to let others do the research. For others, research is

viewed, in varying degrees, as a threat to their product. At one extreme, research is regarded as a debunking enterprise, engaged in by scientists who have little interest in providing human services. At another extreme, the problem is one of educating the researchers in nuance, context, and a clinical approach that emphasizes adapting techniques to changed situations and client tastes. The result is a gap in communication epitomized by two cultures—scientists searching for evidence and practitioners seeking effects and cures. A step toward bridging the gap would consist of mutual education through joint ventures. These ventures would expose scientists to the goals (and motives) of practitioners and would also make practitioners aware of the general analytical approaches used by scientists.

Experimentation is an appropriate vehicle for evaluating performance-enhancing techniques; the problem is usually defined in terms of effects of techniques (procedures) on performance (behaviors). It is also appropriate at an earlier stage in the process, when products are being developed. Products evolve in a kind of trial-and-error fashion similar in many respects to scientific discoveries. One model for integrating research with product development is engineering research and development (R&D). A strenuous applied research effort accompanies the development process in many firms, as does a quality-control program designed to evaluate products both during development and after they have been placed on the market. With a few exceptions, this model has not been adopted by firms or institutions in the field of performance enhancement.

Experimental evidence has accumulated in some areas related to techniques. Although not linked specifically to product development in the manner of an R&D operation, this work does address the question, What evidence supports the claims for the technique? In fact, so strong is the experimental tradition in some areas that a body of work has developed programmatically within a generally accepted paradigm (e.g., guided imagery). The benefits of a long research tradition can be seen in these areas. Meta-analyses have been performed and can be used as a basis for evaluation. For other areas, we are presented with the prospect of relying on scattered experiments or using other criteria as a basis for evaluation, or both (see Appendix A for summaries of the state of the science in each of the areas).

However, the benefits of experimental evidence derive primarily from the general approach rather than from the particular experiments. This idea is captured by Kelman, who noted that "an experimental finding . . . cannot very meaningfully stand by itself. Its contribution to knowledge hinges on the conceptual thinking that has produced it and into which it is subsequently fed back" (1968:161). We emphasize here the contribution

of an analytical approach to thinking about behavior, as distinct from the establishment of laws about psychological processes. It is the cumulation of a series of experiments that winnows out the useful parts of treatments or techniques. It is the self-correcting progression of new experiments that refines treatments, saving those that work and discarding those that do not (or that work only under very restricted conditions). This process contributes equally well to the goals of theory development and product development.

Other evaluation criteria elucidated by Hegge, Tyner, and Genser (1983) include theories, uses, and implications for Army operations and philosophy. A problem with these criteria is that they tend to be vague and somewhat idiosyncratic, making it difficult to propose general categories on which most people would agree. Without precisely defined categories for judging techniques, it is difficult to address issues of transfer of performance from one situation to another or to evaluate newly emerging techniques. A similar problem exists with respect to developing taxonomies in broadly defined fields: there is little agreement on a set of categories for the fields of human learning, performance, motivation, perception, and social and organizational processes. More mature subdisciplines provide an empirical basis for taxonomies, allowing for more tightly constructed systems of tasks and situations: for example, rote learning, short-term memory, concept learning, problem solving, work motivation, and team functions (see Fleishman and Quaintance, 1984). An advantage of such systems is that they capture rather precise relationships between task and performance.

This discussion serves only to introduce the issues and identifies several themes that receive more detailed attention in the chapters to follow. First, any evaluation must take into account the status of the available evidence. Confidence placed in judgments about a technique should be based on the quality of the evidence produced by researchers. Second, the evaluator cannot afford to rely exclusively on a single criterion for judging effectiveness. Theoretical and applied issues are also important, as are considerations of values served or violated by use of the technique. Third, technique development issues are not isolated from research or analytical issues. Each step in the process of product design can be regarded as an empirical issue; decisions made about procedures and packaging can be the result of experimental outcomes. Fourth, the subject of enhancing human performance is not new. It has been a topic of interest for centuries and an area of scientific work for several decades. The literatures on learning and skill acquisition should be consulted by developers, and insights derived from these literatures should be used in product design.

These themes are woven throughout the discussions of specific techniques. Each chapter discusses relevant literature, describes the specific techniques, points to directions for further research when appropriate, and notes possible applications in military and industrial settings. Despite the common coverage, however, each chapter is also unique in that each is tailored to the particular problems associated with its focus.

2

Findings and Conclusions

The committee's first major task was to evaluate the existing scientific evidence for a wide range of techniques that have been proposed to enhance human performance. This evaluation was intended by our Army sponsors to suggest guidelines for decision making on Army research and training programs. In our evaluation we draw conclusions with respect to whether more basic or applied research is warranted, whether training programs could benefit from new findings or procedures, and what, in particular, might be worth monitoring for potential breakthroughs of use to the Army. In many of the areas examined it appears feasible to pursue carefully designed programs that build on basic research; however, such programs should be monitored closely.

The committee's second major task was to develop general guidelines for evaluating newly proposed techniques and their potential application. We are aware that the use of basic and applied research in decision making is a complex issue. Although payoffs from basic research can often be realized in the long run, the value of research findings to the Army depends on developing a way of putting them into practice. With regard to applied or evaluation research, further complexities are evident: multiple, sometimes conflicting, criteria must be satisfied at each of several stages in the evaluation process, from assessing a pilot program to implementing the program in an appropriate setting. Another problem is that of choosing among alternative techniques when none of them has been subjected to a systematic evaluation. In the absence of evaluation studies, the Army needs guidelines for selecting packages and vendors.

The committee's evaluation has produced several answers to questions

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

of how best to improve performance in specific areas. On the positive side, we learned about the possibilities of priming future learning by presenting material during certain stages of sleep, of improving learning by integrating certain instructional elements, of improving skilled performance through certain combinations of mental and physical practice, of reducing stress by providing information that increases the sense of control, of exerting influence by employing certain communication strategies, and of maximizing group performance by taking advantage of organizational cultures to transmit values. On the negative side, we discovered a lack of supporting evidence for such techniques as visual training exercises as enhancers of performance, hemispheric synchronization, and neurolinguistic programming; a lack of scientific justification for the parapsychological phenomena considered; some potentially negative effects of group cohesion; and ambiguous evidence for the effectiveness of the suggestive accelerative learning package.

The remainder of this chapter presents the committee's findings and conclusions, which are presented in two parts: general conclusions regarding the process of evaluating any technique being considered by the Army and specific findings and conclusions for each of the areas of human performance examined. Whenever appropriate, we make recommendations for research, evaluation, and practice.

GENERAL CONCLUSIONS

The committee suggests that the Army move vigorously, yet carefully and systematically, to implement techniques that can be shown to enhance performance in military settings. Such an effort would be timely because of recent developments in the relevant research areas. Moreover, the payoff is likely to be very high if techniques are selected judiciously. Although the desire for dramatic improvements in performance makes some extraordinary techniques attractive, techniques drawn from mainstream research in relevant areas of performance may be more effective. The Army's concern for enhancing human performance and its substantial resources for evaluating techniques place it in a favorable position to take advantage of developments. The Army might also consider the possibilities of transferring its findings to the civilian sector.

Collectively, the committee's conclusions call for the adoption of scientifically sound evaluation procedures; however, these procedures must be adapted to institutional needs and must take into account problems of implementation. We summarize these considerations below.

Scientific Evidence

Techniques and commercial packages proposed for consideration by the Army should be shown to be effective by adequate scientific evidence

or compelling theoretical argument, or both. A technique's utility should be judged in relation to alternatives designed for similar purposes, and the estimated utility should be of significant magnitude. Specific stages of analysis can be incorporated in pilot or field testing, and such testing should be carried out by investigators who are independent of the technique's originators or promoters.

Testimonials as Evidence

Personal experiences and testimonials cited on behalf of a technique are not regarded as an acceptable alternative to rigorous scientific evidence. Even when they have high face validity, such personal beliefs are not trustworthy as evidence. They often fail to consider the full range of factors that may be responsible for an observed effect. Personal versions of reality, which are essentially private, are especially antithetical to science, which is a fundamentally public enterprise. Of course, a caution about testimonials should not be confused with a lack of openness to new and unusual ideas. Such openness is consistent with the requirement that the evidential criteria of science be satisfied.

The subject of testimonials as evidence has received considerable attention in recent research on how people arrive at their beliefs. These studies indicate that many sources of bias operate and that they can lead to personal knowledge that is invalid despite its often being associated with high levels of conviction. The committee recommends that this research be disseminated, as appropriate, in the Army. It may then be applied whenever testimony is used as the primary evidence to promote an enhancement technique.

Conditions for Implementation

Two kinds of evidence should be sought to support decisions to implement a technique: successful field tests and an analysis of implementability. It would also be useful to analyze the impact of the technique or package on the larger system in which it is to be embedded. These analyses would aid in explaining why the procedures are necessary and why certain consequences are expected. In general, any description of what a technique accomplishes should be accompanied by an explanation of *why* it accomplishes what it does. Such an explanation would provide a more fundamental understanding of processes affected by exposure to the technique and permit optimal implementation.

Rational Decision Making

The considerations that must be entertained in selecting a technique for practical use in a military setting are different from the considerations

needed to verify the existence of an enhancement effect in a scientific setting. For example, the benefits of correct decisions and the costs of incorrect decisions, that is, the risk calculus, may differ in the two settings. Furthermore, what is viewed as a timely decision will also differ. The specific differences as they apply to particular decisions should be made explicit.

Mechanisms for Advice

It would be useful to provide valid information about useful techniques to Army commanders and other interested staff on a regular basis. Special consideration should be given to ways in which technique-related information can be transferred from scientists to practitioners. The characteristics of a transfer agent could be defined, and such a position might be established within an appropriate office.

The committee recommends that the Army Research Institute formalize the ways in which it receives and provides advice about specific techniques. A committee to review experimental designs and statistical analyses could be convened to improve the evaluation of techniques. Special and standing committees could also be used to make program recommendations and to review proposals for intramural and extramural research.

Bidding Procedures

Purchase by the Army of a commercial enhancement package should take place within the context of a set of well-defined procedures. The committee recommends that an open-bid procedure be followed, based on a full presentation of the Army's stated objectives. This would encourage competitive evaluation of techniques. The following information, presented in a standard format, should be required: the objectives of the technique, a description of its procedures, evidence that it produces the claimed effects, and the vendor's record of past achievements in relevant areas.

Lack of professional training and research experience in human performance by a designer or advocate should not preclude consideration of the proposed package; it should, however, signal the need for a more stringent analysis by the Army.

SPECIFIC FINDINGS AND CONCLUSIONS

We present below findings and conclusions for each of the areas investigated. Some statements take the form of suggested actions based on what we know; others consist of suggestions for more work or for research that has not yet been done.

Learning During Sleep

1. The committee finds no evidence to suggest that learning occurs during verified sleep (confirmed as such by electrical recordings of brain activity). However, waking perception and interpretation of verbal material could well be altered by presenting that material during the lighter stages of sleep. We conclude that the existence and degree of learning and recall of materials presented during sleep should be examined again as a basic research problem.

2. Pending further research results, the committee concludes that possible Army applications of learning during sleep deserve a second look. Findings that suggest the possibility of state-dependent learning and retention (i.e., better recall of material when learned in the same physiological and mental state) may be applicable to fatigued soldiers. Furthermore, even presentations of material that disrupt normal sleep may be cost-effective, as may presentations that coincide with stages of light sleep.

Accelerated Learning

1. Many studies have found that effective instruction is the result of such factors as the quality of instruction, practice or study time, motivation of the learner, and the matching of the training regimen to the job demands. Programs that integrate all these factors would be desirable. We recommend that the Army examine the costs, effectiveness, and longevity of training benefits to be derived from such programs and compare them with established Army procedures.

2. The committee finds little scientific evidence that so-called superlearning programs, such as Suggestive Accelerative Learning and Teaching Techniques, derive their instructional benefits from elements outside the mainstream of research and practice. We observe, however, that these programs do integrate well-known instructional, motivational, and practice elements in a manner that is generally not present in most scientific studies.

3. We find that scientifically supported procedures for enhancing skills are not being sufficiently used in training programs and make two recommendations to remedy this problem. First, the basic research literature should be monitored to identify procedures verified by laboratory tests to increase instructional effectiveness. Second, additional basic

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

research should be supported to expand the understanding of skill acquisition for both noncombat and combat activities.

4. We conclude that the Army training system provides a unique opportunity for cohort testing of training regimens. The Army is in a position to create laboratory classroom environments in which competing training procedures can be scientifically evaluated.

5. The committee recommends that the Army investigate expert teacher programs by identifying and evaluating particularly effective programs within the Army. In addition, transferable elements of effective instruction can be reported to the larger instructional community.

Improving Motor Skills

1. The committee concludes that mental practice is effective in enhancing the performance of motor skills. This conclusion suggests further work in two directions: (1) evaluation studies of motor skills used in the Army and (2) research designed to determine the combination of mental and physical practice that, on average, would best enhance skill acquisition and maintenance, taking into account both time and cost.

2. The committee concludes that programs purporting to enhance cognitive and behavioral skills by improving visual concentration have not been shown to be effective to date. In our judgment, these programs are not worth further evaluation at this time.

3. The committee concludes that existing data do not establish the generality of observed effects from programs that train visual capabilities to increase performance.

4. Similarly, the committee concludes that the effects of biofeedback on skilled performance remain to be determined.

5. The committee recommends additional research to establish the potential of these techniques in the domain of specific skilled performances.

Altering Mental States

1. Time did not allow the committee to explore the evidence for a wide variety of specific methods for relating mental states to changes in performance. Such methods include forms of self-induced hypnotic states and peak performance resulting from high levels of focused concentration and meditation. We recommend that reviews of the literature in these areas be undertaken to ascertain whether any practical results might be obtained by the use of such methods.

2. The committee finds that, while the study of mental computations in language and imagery has progressed in recent years, the effort to understand how such computations are modulated by energetic factors

such as arousal, stress, emotion, and high levels of sustained concentration has not been fully developed. For example, the claims that certain mental states produce general improvements in performance derive from the idea, supported by research, that arousal affects mental computations and that there ought to be an optimal level of arousal for the performance of such computations. We recommend this as an important area for investment of basic research funds.

3. The committee's review of the appropriate literature refutes claims that link differential use of the brain hemispheres to performance. Further evaluation of these claims depends on developing valid and reliable measures of hemispheric involvement.

4. The committee finds no scientifically acceptable evidence to support the claimed effects of techniques intended to integrate hemispheric activity, for example, Hemi-Sync™. Attempts to increase information-processing capacity by presenting material separately to the two hemispheres do not appear to be useful. We conclude that such techniques should be considered further by the Army only if scientific evidence is provided to and evaluated by the Army Research Institute.

Stress Management

1. Existing data indicate that stress is reduced by giving an individual as much knowledge and understanding as possible regarding future events. In addition, giving the individual a sense of control is effective. On the basis of these findings, the committee recommends a systematic program of research and development that would address three questions: (1) How relevant is this finding for stress reduction in the Army? (2) To what extent does stress reduction realized in training transfer to combat situations? (3) What are the limitations on providing knowledge and understanding of future events and a sense of control in the Army setting? Pending the outcome of this research, we suggest that consideration be given to including the material in training programs for company grade, field grade, command, and staff officers.

2. We find that, while biofeedback can achieve a reduction of muscle tension, it does not reduce stress effectively. It is therefore not a promising research topic in that respect. We recommend that funding be directed toward investigation of more promising stress management procedures.

3. We recommend that information be gathered on the costs of stress in terms of organ breakdown, loss of efficiency, and loss of time. This information would have implications for training programs.

Influence Strategies

1. The committee finds no scientific evidence to support the claim that neurolinguistic programming is an effective strategy for exerting influence.

We advise that further Army study of this aspect of NLP be made only in comparison with other techniques.

2. There are no existing evaluations of NLP as a model of expert performance. We conclude that further investigation of such models may be worthwhile and suggest that NLP be examined in comparison with several other techniques.

3. Concerning the process of technology transfer, we recommend that studies be conducted to develop training regimens for those who train others to wield social influence. The large literature on this topic in social psychology would provide a basis for such packages.

Group Cohesion

1. We find few scientific studies that address the possible relationship between group cohesion and performance; however, such a relationship may well be found with more extensive research. There is a need for research to consider the possibility of negative effects from inducing cohesion and methods of avoiding such effects. The committee recommends continued study of cohesion and related group processes.

2. We are favorably impressed with the evaluation studies of the Army's COHORT system. We endorse the investigators' plan to proceed beyond measures of attitudes to measures of group performance.

3. We recommend that the Army, as well as independent investigators, study the possible impacts of cohesion beyond the COHORT system, for example, on intergroup performance.

Parapsychology

1. The committee finds no scientific justification from research conducted over a period of 130 years for the existence of parapsychological phenomena. It therefore concludes that there is no reason for direct involvement by the Army at this time. We do recommend, however, that research in certain areas be monitored, including work by the Soviets and the best work in the United States. The latter includes that being done at Princeton University by Robert Jahn; at Maimonides Medical Center in Brooklyn by Charles Honorton, now in Princeton; at San Antonio by Helmut Schmidt; and at the Stanford Research Institute by Edward May. Monitoring could be enhanced by site visits and by expert advice from both proponents and skeptics. The research areas included would be psychokinesis with random event generators and Ganzfeld effects.

2. One possible result of the monitoring mentioned above is the proposal

of specific studies. In that situation the committee recommends the following procedures: first, the Army and outside scientists should arrive at a common protocol; second, the research should be conducted according to that protocol by both proponents and skeptics; and third, attention should be given in such research to the manipulability and practical application of any effects found to exist.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

3

Evaluation Issues

Implementation of an enhancement technique, in the committee's view, should depend on two general kinds, or levels, of evaluation. The first examines primarily the scientific justification for the effectiveness of the technique and the potential of the technique for improving performance in practice. The second kind examines field tests of a pilot program incorporating the technique to determine how feasible it is and to what extent it brings about effects that Army officials consider useful.

Convincing scientific justification can come only from basic research, that is, from carefully controlled studies that usually take place in laboratory settings and that preferably are related to a body of theory. Such research can provide evidence for the existence of the causal effect on which a technique is based and can help explain, or indicate a mechanism for, the effect. Analysis in connection with basic research should go beyond scientific justification to operational potential and likely cost-effectiveness. Only field tests can assess a program's actual operations and effects, however, and for such tests a broader array of evaluative criteria are needed, related primarily to the technique's utility.

Because strong claims of support from basic research have been made for some of the techniques the committee examined, we review here what it takes to justify a scientific claim, specifically, we review some standards for evaluating basic research. We then examine in more detail some standards for evaluating field tests of pilot programs. In the third section of this chapter, we set forth briefly some of our impressions of how the Army now manages the solicitation and evaluation of new performance-enhancing techniques. This chapter concludes with a note

on informal, qualitative approaches to evaluation, which are sometimes suggested as alternatives to basic research and field tests.

This chapter does not aspire to a comprehensive treatment of evaluation issues, and it barely touches on research methods. Articles, journals, books, and handbooks testify to the scope and complexity of this burgeoning field (e.g., Barber, 1976; Cook and Campbell, 1979). Our objective here is to highlight the topics that have impressed us as most germane. The various sources just mentioned would need to be consulted for even a minimal elaboration of these topics, and other committees would be required if recipes for evaluation of the Army's enhancement programs were sought as extensions of our work. Still, we believe this chapter will help the Army set general evaluation standards.

STANDARDS FOR EVALUATING BASIC RESEARCH

The purpose of basic research is to permit inferences to be drawn in accordance with scientific standards, including inferences about novel concepts, about causation, about alternative explanations of causal relations, and about the generalizability of causal relations.

For novel concepts, evidence must be gathered that both the purported enhancement technique and the relevant performance have been (1) defined in a way to highlight their critical elements, (2) differentiated from related variables that might bring about similar effects, and (3) put into operation (manipulated or measured) in ways that include the critical parts. The burden is on the evaluator to analyze how the components of each new technique differ from concepts already in the literature. The need for this standard is illustrated well by packages for accelerated learning, as discussed in Chapter 4.

Evidence needs also to be adduced that supposed cause and effect variables vary together in a systematic manner. Relevant procedures include comparison of performance before and after introduction of the technique, contrasts of experimental and control groups in an experimental design, and calculation of statistical significance. Illusory covariation can occur more easily in nonstatistical studies, which are used often to support the existence of paranormal effects, as discussed in Chapter 9.

Especially demanding is the need for evidence that the performance effect observed is due to the postulated cause and not to some other variable. Ruling out alternative explanations or mechanisms requires intimate knowledge of a research area. Historical findings and critical commentary are needed to identify alternatives, determine their plausibility, and judge how well they have been ruled out in particular sets of experiments. Common threats to the validity of any presumed cause-

effect relation include effects stemming from subject selection, unexpected changes in organizational forces, the spontaneous maturation of subjects, and the sensitizing effects of a pretest measurement on a posttest assessment. Experiments with random assignment of subjects to treatments are preferred, but some of the better quasi-experimental designs are also useful. Another class of threats to validity is associated with subject reactions to such conceptual irrelevancies as experimenter expectations about how subjects should perform or subjects' performing better merely because they are receiving attention. Procedures that have evolved to reduce this sort of threat include double-blind experiments, placebo control groups, mechanical delivery of treatments, and the elimination of all communication between experimenters and subjects or among subjects. These safeguards, however, are not certain, and implementing them is not a simple matter.

Finally, for a technique to be of value, one must ascertain that a causal relation observed in one setting is likely to be observed in other settings in which the technique is to be employed. Replication of an experiment by an independent investigator is a first step. Another step is to produce the cause and effect with different samples of people, settings, and times. Systematic reviews of the literature, perhaps aided by what is referred to as meta-analysis of studies (as illustrated in Chapter 5), are also helpful. Beyond these steps, a thorough theoretical understanding of causal processes, which is a fundamental goal of science, permits increased practical control.

Our point—perhaps seeming obvious to many but nonetheless needing emphasis here—is that a planned or existing program for implementing an enhancement technique is much more likely to bear fruit if evidence for the technique's effectiveness is properly derived from basic research. A complex set of ground rules exists for conducting and drawing inferences from basic research, and waiving those rules greatly increases the chances of incorrect conclusions.

STANDARDS FOR EVALUATING FIELD TESTS OF PROGRAMS

An adequate appraisal of an actual enhancement program requires attention to three general factors. First, the organizational (i.e., political, administrative) context in which the program is embedded should be described. That context strongly influences the choice of evaluation criteria, the types of evaluations considered feasible, and the extent to which evaluation results will be used. Second, the program's consequences should be described and explained, including planned and unplanned, short-term and long-term consequences. The way the program

is construed influences the claims resulting from an evaluation and the degree of confidence that can be placed in what was learned. Third, value or merit should be explicitly assigned to a program. Valuing relates an enhancement technique to an Army need and to feasible alternatives. In the following sections we comment on these three factors in turn.

The Organizational Context

A description of the broader context of an enhancement program would include an assessment both of the various constituencies with a stake in its implementation and of the priorities of the larger institution. We do not discuss stakeholder interests in general at this point because we refer to some specifically later in this chapter, in the section on the committee's impressions of current Army evaluation practices. We do comment here on the Army's institutional priorities as they may relate to scientific standards.

We understand that the Army, like other organizations in society, may have—and quite possibly should have—different standards for evaluating knowledge claims, or technique effectiveness, than science has. The scientific establishment is conservative in the tests it administers to discipline its conjectures; in particular, its goal is to reduce uncertainty as far as possible, no matter how long that takes. In the Army, by contrast, the need for timely information and decisions may lead to an acceptance of greater uncertainty and a higher risk of being wrong.

There is no Army doctrine of which we are aware concerning the degree of risk that is acceptable in evaluations of pilot programs. Yet surely one objective of evaluations of pilot programs should be to describe the costs to the Army of drawing incorrect conclusions so that inferential standards can be made commensurate with those costs. If the costs are relatively low, the riskier approach of most commercial research (as, for example, in management consulting or marketing) may be preferred to the more conservative approach of basic science.

Describing a Program's Consequences

In evaluating a program, it is desirable to present an analysis and defense of the questions probed and not probed, together with justification for the priorities accorded to various issues. Primary issues usually include the program's immediate effects and its organizational side effects.

Immediate Effects

A primary problem in evaluation is to decide on the criteria by which a program is to be assessed. The major sources for identifying potential

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

criteria include program goals, interviews with interested persons, consideration of plausible consequences found in the literature, and insights gained from preliminary field work.

Such criteria specify only potential effects, however. They do not speak to the matter of whether the relation between a supposed cause and effect is truly causal. In this respect, a fundamental issue of methodology is the use of randomized experiments. Although logistic reasons abound in any practical context for not going to the trouble to use such research designs, one might nonetheless argue that the Army is in a better position to conduct randomized experiments than are organizations in such fields as education, job training, and public health. The reason for going to such trouble is that randomized experiments give a lower risk of incorrect causal conclusions than the alternatives.

Alternatives at the next level of confidence are quasi-experimental designs that include pretest measures and comparison (control) groups. Relatively little confidence can be placed either in before-after measurements of a single group exposed to a technique without an external comparison, or in comparisons of nonequivalent intact groups for which pretest measures are not available.

Side Effects

Unintended side effects include impacts on the broader organization, and these should be monitored. For example, trainers from other (nonexperimental) units may copy what they think is going on, or they may simply be upset by the implementation of new instructional packages in the experimental units. Units not treated in the same way as the experimental units may be unwilling to cooperate when cooperation would seem to be in their best interest. They may also suffer by comparison, as is thought to be the case, for example, when COHORT units are introduced into a division (see Chapter 8). Evaluators should strive to see any program as fitting into a wider system of Army activities on which it may have unintended positive or negative effects.

Assigning Value to Pilot Programs

The described consequences of a program tell us what a program has achieved but not how valuable it is. Three other factors are important in inferring value: Does the new technique meet a demonstrable Army need to the extent that without it the organization would be less effective? How likely is it that the program can be transferred to other Army settings, either as a total package or in part? How well does the new

program fare when compared with current practice and with alternatives for bringing about the same results?

Meeting Needs

Representatives of the commercial world who seek outlets for their products often confound wants with needs, enthusiasm with proof, and hope with reality. While it is axiomatic that all field tests should aim to meet genuine Army needs, it is not clear how needs are now assessed when the developers of new products approach Army personnel for permission to do general research or field tests. It is clear that a needs analysis should be part of the documentation about every field test.

What should a needs analysis look like? At the minimum, it should document the current level of performance at some task, why the level is inadequate, what reason there is to believe that performance can change, and what the Armywide impacts would probably be if the performance in question were improved. In addition, an analysis should question why a particular program is needed for solving the problem. Such an analysis would describe the program, critically examine its justification in basic research, identify the financial and human resources required to make the program work, relate the resources required to the funds available, examine other ways of bringing about the same intended results, and justify the program at hand in terms of its anticipated cost-effectiveness. To facilitate critical feedback, such reports should be independent of the persons who sponsor a program, though based on a thorough, firsthand acquaintance with the program and its developers and sponsors.

As just described, needs analysis is a planning exercise to justify mounting a pilot program. It is not a review of program achievements relative to needs, for which a description of a program's consequences is required. At that later stage in evaluation a judgment is required about whether the magnitude of a program's effects is sufficient to reduce needs to a degree that makes a practical difference. More is at stake than whether the program makes a statistically reliable difference in performance. Size of effect relative to need is the crucial concern. When the magnitude of change required for practical significance has been specified in advance, it is easy to use such a specification to probe how well a need has been met. But the level of change required to alleviate need is not usually predetermined, and there are political reasons why developers are not always eager to have their programs evaluated in terms of effect sizes they themselves have clearly promised or that others have set for them.

Needs can be specified only by Army officials, and it is vital that such

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

officials inspect the results a program has achieved, relating them to their perception of need. Since the Army is heterogeneous, it would be naive to believe that there are no significant differences within it about how important various needs are and how far a particular effect goes in meeting a particular need. Some theorists relate needs primarily to the number of persons performing below a desired level, while others emphasize the seriousness of consequences for unit performance, for which deficiencies in only one or two persons may be crucial. Some practitioners are likely to think a deficit in skill X is worse than a deficit in skill Y, while others may believe the opposite. Evaluators who take the concept of need seriously have to take cognizance of such heterogeneity, perhaps using group approaches like the Delphi technique to bring about consensus on both the level of need and the extent to which a particular pattern of evaluative results helps meet that need.

Likelihood of Transfer

Although some local commanders may sponsor field trials for the benefit of their command alone, the more widely a successful new practice can be implemented within the Army, the more important it is likely to be. Consequently, evaluations of pilot programs should seek to draw conclusions about the likelihood that findings will transfer to populations and settings different from those studied.

In this regard, it is particularly important to probe the extent to which any findings from a pilot study might depend on the special knowledge and enthusiasm of those persons who deliver or sponsor the program. Such persons are often strongly committed to a program, treating it with a concern and intensity that most regular Army personnel could not be expected to match. While it is sometimes possible to transfer such committed persons from one Army site to another in order to implement a program, in many instances this cannot be done. Transfer is partly a question of the psychology of ownership; authorities who did not sponsor a product will sometimes reject out of hand what others have developed, including their immediate predecessors. Since Army leaders in any position turn over with some regularity due to transfers, promotions, and retirement, successors will probably not identify with a program as strongly as the original sponsors and developers did.

The likelihood of transfer also affects the degree to which program implementation is monitored. Pilot programs are likely to be more obtrusively monitored than other programs. Not only is this obtrusiveness due to developers' and evaluators' fussing over their charge, it is also due to teams of experts brought in to inspect what is novel and to responsible officers wanting to show others the unique programs they

are leading (and on which the success of their careers may depend). For at least these reasons pilot programs tend to stand out more than the regular programs they may engender. Research suggests that the quality with which programs are delivered may in fact increase when outside personnel are obviously monitoring individual and group performance.

It is naive to believe that one can go confidently from a single pilot program to full-blown Armywide implementation. Even if this were feasible politically, it would not be technically advisable unless there were compelling evidence from a great deal of prior research indicating that the program was indeed built on valid substantive foundations. Given a single pilot program, decisions about transfer are best made if the program is tested again, at a larger but still restricted set of sites and under conditions that more closely approximate those that would pertain if the new enhancement technique were implemented as routine policy. Only then might serious plans for Armywide implementation be feasible.

Contrast with Alternatives

Most of the evaluation we have discussed contrasts a novel program with standard practices that are believed worth improving; yet rational models of decision making are usually predicated on managers' having to choose among several different options for performing a particular task. One would hope that every sponsor of a novel performance enhancement technique is conversant with the practical alternatives to it and has cogent arguments for rejecting them.

Many novel techniques have some components that are already in standard practice or can be clearly derived from established theories. Upon close inspection, pilot programs often turn out to be less novel than their developers and sponsors claim. Of course, the Army may often find it convenient to order complete packages in the form offered and may not have much latitude to interact with developers in order to modify package contents to emphasize what is truly a novel alternative and to downplay that which is merely standard practice.

Ultimately, alternatives have to do with costs. Although many forms of cost are at issue—including those associated with how much a new practice disrupts normal Army activities and how much stress it puts on personnel—the major cost usually considered is financial. Cost analysis is always difficult, nowhere more so than in the Army, which uses many ways to calculate personnel costs. Nonetheless, in planning an evaluation, some evidence about the total cost of a pilot program to the Army will usually be available and can be critically scrutinized. It is also useful, as far as possible, to ascribe accurate Army costs to each of the major components of such an intervention. In our view, what is called cost-

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

effectiveness analysis lends itself better than what is called cost-benefit analysis to the comparison of different programs. The purpose of cost-effectiveness research is to express the total cost for each program in dollar terms and to relate this to the amount of effect as expressed in its original metrics—unlike cost-benefit research, in which even the effects have to be expressed in dollar terms. Sophisticated consumers of evaluation should want something akin to cost-effectiveness knowledge, for it reflects decisions they should be making. Is it not useful to know, for example, that the best available computer-assisted instruction packages are much less cost-effective than peer tutoring?

CURRENT STATUS OF ARMY EVALUATIONS

We set forth here some of our impressions of the way in which the Army currently manages the solicitation and evaluation of novel techniques to enhance performance. We must stress that these are only impressions, gained through the limited investigative capabilities of a committee such as ours, not hard conclusions based on systematic research directed at the particular question. Furthermore, although the opinions that follow are largely critical of Army procedures, they are not accompanied by much detail. As noted earlier, the focus here is on the identification of the various Army constituencies that have a stake in enhancement programs and on the role they play in evaluation.

How the Army decides which among competing proposals should be sponsored for development or for field tests is not clear. What is clear is that decision making is diffuse both geographically and institutionally. Sponsorship may come from senior managers in the Pentagon or from local personnel of varying rank. While differences in the quality of program design, implementation, or evaluation may be correlated with the source of sponsorship, such a correlation is not clear at present in the Army context.

A particular concern is that Army sponsors of pilot programs may base their judgment about the value of a program either on their own ideas about what is desirable or effective or on the persuasiveness of the arguments presented to them by program developers, who stand to gain financially if the Army adopts their program. Judgments of value should depend on broader analysis of Army needs and resources, as well as on realistic assessment of the quality of proposed ideas based on a thorough and independent knowledge of the relevant research literatures. Sponsors should examine what is being advocated at every stage: proposal, testing, and implementation.

Also of concern when pilot programs are planned is how decisions are reached about funding and about the quality of implementation expected

from them. Although systematic evidence is lacking, it seemed to committee members that pilot programs are not generally implemented well and, except for fiscal accountability, are not closely monitored by their Army sponsors. Evaluations of pilot programs should try to characterize resources required by the program and the resources actually available.

We found little evidence that sponsors, advocates, or local implementers had aspirations to evaluations that use state-of-the-art methods. We found no guidelines about the standards expected for evaluative work, whether in the form of published minimal standards or published statements of preferred practices. When it comes to field trials of novel ideas for enhancing human performance, the monitoring of evaluation quality does not seem to be part of the organizational context. Given the absence of formal expectations in these regards, it is not surprising that the pilot programs we saw and the evaluation materials we read were usually disappointing in the technical quality of the research conducted. In settings in which program sponsors or advocates control an evaluation, weaker evaluations (e.g., based on testimony) will sometimes be preferred to stronger methods (e.g., experiments) because the latter are usually more disruptive when implemented and are more likely to result in effects that are disappointing, however much more accurate they may be. The weaker methods are easier to implement when few units are available, are less disruptive of ongoing activities, are easier to manipulate for self-interested ends, and need not be as expensive for data collection.

We saw little evidence that the Army requires evaluations by persons independent of the pilot program under review. Moreover, the nonindependent evaluations we saw did not seem to have been subjected to any of the peer review procedures to which research results (and plans) are subjected not only in academic sciences, but also in much of the corporate world, as with, say, pharmaceutical testing. While in-house evaluation is highly valuable for gaining feedback for program improvement, many experienced evaluators contend that it is inadequate for assigning overall value because in-house evaluators cannot divorce themselves from their own stake in the program under examination. Although it is not easy to specify organizational standards adequate for a high-quality field test of some novel technique, it is also not difficult to detect the inadequacies associated with local program sponsors' having few clear expectations about the desirable qualities of program operations or evaluative practices. In the absence of such expectations, program developers and evaluators may believe that few officials care about the small-scale field tests of techniques on which the developers'—and, all too often, the evaluators'—own welfare depends.

Since the organizational climate we have just described is not optimal

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

for gaining trustworthy information about program value, future evaluators of Army field trials might do well to characterize: (1) what program managers expect in terms of the quality of the program and its evaluation; (2) who is paying attention to the trials; and (3) for what purposes they want to use any information provided by the evaluation. This kind of information, as mentioned above, contributes to a description of the organizational context of a program, which is a major part of an adequate evaluation.

QUALITATIVE APPROACHES

Alternatives to experimentation are the largely qualitative traditions, which rely mostly on direct observation, sometimes supplemented by archival data. Investigative journalists operate in this mode; so do many cultural anthropologists, political scientists, and historians. These professions use clues to suggest hypotheses about possible causes and investigate the empirical evidence in ever-greater detail in an attempt to rule out hypotheses until they are left with just one. A critical aspect of their work is the use of substantive theories and ad hoc findings from the past to help in ruling out alternative explanations. Also working in this tradition are committees of psychologists who seek to make statements about the causes of enhanced human performance. Rarely conducting studies themselves, they instead sift through historical evidence provided by reviews of the literature and make on-site observations in the manner of detectives, pathologists, investigative journalists, and cultural anthropologists.

These traditions rely strongly on personal testimony. Respondents' reports are taken seriously and, indeed, should be. Any method can, in principle, generate strong causal evidence, provided that plausible alternatives to a preferred hypothesis have been ruled out. The general issues are: Can personal testimony usually rule out all the plausible alternative interpretations? Does use of it engender the very threats to validity that militate against strong inferences? Dale Griffin, in a paper prepared for the committee (see Appendix B), suggests "no" to the first question and "yes" to the second. His analysis of biases that operate when people attempt to explain how and why they changed after an experience reveals many of the shortcomings associated with relying on testimony as a major means of testing causal hypotheses.

While testimony can be regarded as a form of confirmatory evidence, it does not provide any of the disconfirming evidence needed to reduce uncertainty. Rarely are there the kinds of comprehensive probes needed to discover why respondents believe that the effects are due to a treatment rather than to maturation, statistical regression, or the pleasant feelings

aroused by the experiences. People are typically weak at identifying the range of such alternatives, however simply they may be described, and at distinguishing the different ways in which the causal forces might operate. How can people know how they would have matured over time in the absence of an intervention (technique) that is being assessed? How can people disentangle effects due to a pleasant experience, a dynamic leader, or a sense of doing something important from effects due to the critical components of the treatment per se? Much research has shown that individuals are poor intuitive scientists and that they recreate a set of known cognitive biases (Nisbett and Ross, 1980; Griffin). These include belief perseverance, selective memory, errors of attribution, and overconfidence. These biases influence experts and nonexperts alike, usually without one's awareness of them. Scientists hold these biases in partial check by using random assignment instead of testimony and by the tradition of public scrutiny to identify and analyze alternative interpretations for observed events. Such methodological traditions can be transmitted to consumers and producers of enhancement techniques through courses on statistical inference and formal decision making. These courses would have the salutary effect of calling attention to the shortcomings of testimony as evidence.

We submit that experimental methods facilitate causal inferences better than the alternatives. They reduce more uncertainty by ruling out more of the contending interpretations for observed effects. However, we refer here to the *relative* superiority of experimentation; such superiority should not be confused with either the perfection or even the adequacy of experimentation. Its problems include the facts that experiments cannot be implemented under all conditions and that experimentation has its own set of unintended side effects. Thus, experimental methods do not guarantee causal inferences and so cannot obviate the need for critical analysis that, on a case-by-case basis, is sensitive to the contexts and traditions of particular institutions or communities, such as the Army, on one hand, and the various promoters of new enhancement techniques, on the other. Moreover, well-conceived research is costly: it requires specially trained investigators, equipped facilities, and programs that may need extensive collaborations and review panels. It is also a demanding craft that requires sensitivity to detail and precision in order to ensure results that are interpretable.

On balance, the benefits derived from careful experimentation outweigh the costs just mentioned. All other things being equal, experimentation is much the preferred strategy for judging the efficacy of techniques that purport to enhance performance, and it should be used whenever possible.

PART III

Parapsychological Techniques

OF ALL THE SUBJECTS TREATED in this volume, none is more controversial than parapsychology. While the flavor of the debates is captured to some extent in this chapter, the subject is treated in the same manner as the other techniques reviewed: we address the question of whether the evidence warrants further consideration of parapsychological techniques for research or application or both.

Emphasized here is information gathering by remote viewing and mind-over-matter effects in controlling machine behavior, particularly machines that generate series of random numbers, which are often used in parapsychology experiments. Although scattered results are said to be statistically significant, an evaluation of a large body of the best available evidence does not support the contention that these phenomena exist. If, however, future experiments, conducted according to the best possible methodological standards, are more generally viewed as producing significant results, it would be appropriate to consider a systematic program of research. Such a program should include a concern for the need to proceed from small effects to practical applications.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

9

Paranormal Phenomena

BACKGROUND

The primary purpose of this chapter is to evaluate the scientific evidence on parapsychological techniques in selected areas. A more complete understanding of the topic, however, requires that we provide background on the military's interest in these phenomena and treat the conceptual issue of how people come to believe as they do. This background section includes a discussion of the phenomena and the military's interest in them as well as an overview of the committee's focus. A brief examination of the different kinds of justifications for the claims is followed by a more detailed treatment of the evidence in areas that have produced large literatures: remote viewing, random number generators, and what are called Ganzfeld (whole visual field) experiments. In addition, we describe experimental work that the committee actually witnessed by visiting a parapsychological laboratory. Despite the growing scientific tradition in some of these areas, many people continue to rely on qualitative or experiential evidence to support their beliefs; we discuss the problems associated with qualitative evidence in conjunction with the research on cognitive and emotional biases, which is reviewed in the paper by Dale Griffin (Appendix B). Finally, the chapter summarizes the committee's major conclusions.

The Nature of the Phenomena

Parapsychologists divide *psi*—the term applied to all psychic phenomena—into two broad categories: *extrasensory perception* (ESP) and

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

psychokinesis (PK). Included in ESP are telepathy, precognition, and clairvoyance, all of which refer to methods of gathering information about objects or thoughts without the intervention of known sensory mechanisms. Popularly called mind over matter, PK refers to the influence of thoughts upon objects without the intervention of known physical processes.

A presentation to the committee by several military officers described in some detail the results of experiments in remote viewing carried out at both SRI International and the Engineering Anomalies Research Laboratory at Princeton University. In these experiments subjects are said to have more or less accurately described a geographical location being visited by a target team. Although the human subjects have no way of normally knowing the target location, the examples recounted appear to indicate, at first glance, some striking correspondences between their descriptions and the actual sites. These studies have been related by some persons to reported out-of-body experiences.

The presentation included discussion of psychic mind-altering techniques, the levitation claims of transcendental meditation groups, psychotronic weapons, psychic metal bending, dowsing, thought photography, and bioenergy transfer. It was indicated that the Soviet Union is far ahead of the United States in developing potential applications of such paranormal phenomena, in particular psychically controlling and influencing minds at a distance. At the presentation, personal accounts were given of spoon-bending parties, in which participants believe they have caused cutlery to bend with the power of their minds, as well as instances of self-hypnosis to control pain and cure illness, walking barefoot on fire and handling hot coals without being burned, leaving one's body at will, and bursting clouds by psychic means.

The media and popular publications, especially in recent years, have discussed various aspects of psychic warfare. Three recent books, by Ebon (1983), McRae (1984), and Targ and Harary (1984), have attempted to document Soviet and American efforts to develop military and intelligence applications of alleged paranormal phenomena. These accounts have been augmented by newspaper stories, magazine articles, and television programs. Many of these sources acknowledge the speculative nature of the proposed applications, but others report that some of the techniques already exist and work.

The claimed phenomena and applications range from the incredible to the outrageously incredible. The "antimissile time warp," for example, is supposed to somehow deflect attack by nuclear warheads so that they will transcend time and explode among the ancient dinosaurs, thereby leaving us unharmed but destroying many dinosaurs (and, presumably, some of our evolutionary ancestors). Other psychotronic weapons, such as the "hyperspatial nuclear howitzer," are claimed to have equally bizarre capabilities. Many of the sources cite the claim that Soviet psychotronic weapons were responsible for the 1976 outbreak of Legionnaires' disease, as well as the 1963 sinking of the nuclear submarine *Thresher*.

Potential Military Applications

Some people, including some military decision makers, can imagine potential military applications of the two broad categories of psychic phenomena. In their view, ESP, if real and controllable, could be used for intelligence gathering and, because it includes "precognition," ESP could also be used to anticipate the actions of an enemy. It is believed that PK, if realizable, might be used to jam enemy computers, prematurely trigger nuclear weapons, and incapacitate weapons and vehicles. More specific applications envisioned involve behavior modification; inducing sickness, disorientation, or even death in a distant enemy; communicating with submarines; planting thoughts in individuals without their knowledge; hypnotizing individuals at a distance; psychotronic weapons of various kinds; psychic shields to protect sensitive information or military installations; and the like. One suggested application is a conception of the "First Earth Battalion," made up of "warrior monks," who will have mastered almost all the techniques under consideration by the committee, including the use of ESP, leaving their bodies at will, levitating, psychic healing, and walking through walls.

The Committee's Focus

Although such colorful examples provide the context for our agenda, the cumulative body of data in the discipline of parapsychology enables us to judge the degree to which paranormal claims should be taken seriously. Since 1882 reports of both naturally occurring incidents and phenomena in laboratory settings have been accumulated in journals, monographs, and books. Just to survey the reports in the refereed journals of parapsychology would be an enormous undertaking. As scientists, our inclination is, of course, to restrict ourselves to the evidence that purports to be scientific. But the alleged phenomena that have apparently gained most attention and that have apparently convinced many proponents do not come from the parapsychological laboratory. Nothing approaching a scientific literature supports the claims for psychotronic weaponry, psychic metal bending, out-of-body experiences, and other potential applications supported by many proponents.

The phenomena are real and important in the minds of proponents, so

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

we attempt to evaluate them fairly. Although we cannot rely solely on a scientific data base to evaluate the claims, their credibility ultimately must stand or fall on the basis of data from scientific research that is subject to adequate control and is potentially replicable.

We divided the task into two parts. First, we looked at the best scientific arguments for the reality of psychic phenomena. Our sponsors, as well as our own appraisal of the current status of parapsychology, indicated that the two most influential scientific programs were the experiments on remote viewing and the experiments on psychokinesis using random event generators. In addition, we looked at the research on the Ganzfeld (whole visual field) because this, in the opinion of many parapsychologists, is the most likely candidate for a replicable experiment. We also report on a parapsychological experiment that the committee itself witnessed.

Second, we considered the arguments of proponents who rely on what they call qualitative as opposed to quantitative evidence for the paranormal. Such evidence depends on personal experience or the testimony of others who have had such experience. Most, if not all, of this evidence cannot be evaluated by scientific standards, yet it has created compelling beliefs among many who have encountered it. Witnessing or having an anomalous experience can be more powerful than large accumulations of quantitative, scientific data as a method of creating and reinforcing beliefs. Because personal experience rather than scientific data has been the source of most beliefs in the paranormal, we have devoted some of our resources to considering this sort of cognitive method as a tool for achieving knowledge.

STANDARDS OF EVIDENCE

Diverse justifications have been offered for pursuing paranormal claims. One argument asserts that paranormal phenomena may no longer be anomalous, given the implications of contemporary quantum mechanics. Indeed, a few physicists have supported some parapsychologists in maintaining that certain forms of precognition and psychokinesis are consistent with some interpretations of quantum theory. The other major argument is that we have no choice but to get involved because the Soviet Union already has a program to develop military applications of psychic phenomena.

Several proponents, including some scientists, firmly believe that paranormal phenomena have been scientifically demonstrated several times over. At the same time, most scientists do not believe that psi exists. Many persons on both sides believe this paradox to be the result of irrational and dogmatic belief systems. The proponents accuse the critics of being closed-minded and bigoted. The critics imply that the

proponents have allowed wishful thinking to bias their judgment and that they are incompetent scientists and are self-deceived. Both sides can point to examples to back their positions.

One essential question confronts the committee: What does an impartial examination of the scientific evidence reveal about the existence of psi? Such an examination assumes that clear standards exist for judging the adequacy of the evidence, which, in turn, raises the issue of what constitutes sufficient evidence. That issue involves many difficult philosophical, theoretical, and methodological matters. For example, Palmer, in his "An Evaluative Report on the Current Status of Parapsychology" (1985), denies that current parapsychological experiments can provide any evidence for the existence of psi. This is because psi implies paranormality and, according to Palmer, we cannot argue that a given effect has a paranormal cause until we have an adequate theory of paranormality. He further argues, however, that parapsychological experiments can and do provide evidence for the existence of anomalies. By an anomaly, Palmer means a statistically significant deviation from chance expectation that cannot readily be explained by existing scientific theories. The burden of Palmer's paper is that just such anomalies have been demonstrated.

Because parapsychologists other than Palmer do not make this distinction between demonstrating an anomaly and testing a theory of paranormality, we do not carry on this distinction in our own assessment of the evidence. We tend to agree with Palmer on this matter, however. When we talk about evidence for psi in the remainder of this chapter, we are using psi in the neutral sense of an apparent anomaly rather than in the stronger sense of a paranormal phenomenon.

Minimal Criteria

Fortunately, critics and parapsychologists appear to agree on the general requirements necessary to demonstrate psi in a parapsychological experiment. Both Palmer (1985) and James E. Alcock (Appendix B) discuss such criteria in their respective papers. As Palmer points out, psi is defined negatively as a statistical departure from a chance baseline that cannot be accounted for by chance, sensory cues, or known artifacts. Such a negative definition implies the minimal criteria required to justify a conclusion that psi has been demonstrated.

Given the statistical aspect, it is imperative that the data be collected in such a way that the underlying probability model and assumptions of the statistical test are fulfilled. This means that targets must be adequately randomized and that each trial in the experiment must be independent of the preceding ones—and, of course, the statistical procedures must be

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

applied and interpreted correctly. Given that all ordinary explanations must be ruled out, the experimenter must take special precautions to ensure that sensory cues, recording errors, subject fraud, and other alternatives have been prevented. Although it is impossible to rule out completely every possible contaminant or to anticipate every alternative, there are reasonable standards that most parapsychologists would agree should be followed.

Because different research paradigms have their own special requirements, no single set of standards can be specified in advance for all parapsychological experiments. Experiments with electronic number generators, for example, rarely have problems with data recording, but they do require special methods such as tests of randomness and attention to the immediate physical environment that are unnecessary with more traditional parapsychological experiments. One requirement for assessing the adequacy of a given experiment is that its procedures and methods of analysis be adequately documented. Unless we know how the targets were selected, how the results were analyzed, how the possibility of sensory leakage was prevented, and how other such aspects of the study were carried out, we have no basis for evaluating the quality of the information provided by the experiment.

Global Criteria

The criteria mentioned in the preceding paragraphs apply to the individual experiment. More global criteria come into play when one wants to evaluate an entire research program or set of experiments. Here we look for such things as replicability, robustness, lawfulness, manipulability, and coherent theory. These criteria deal with the coherence and intelligibility of the alleged phenomena. It is in terms of such global criteria that parapsychological research has been especially vulnerable.

Much of the objectivity involved in assessing the adequacy of research applies to judging individual experiments. But science is cumulative and depends not so much on the outcome of a single experiment as on consistent and lawful patterns of results across many experiments carried out in a variety of independent settings. Lawful consistency in this sense, according to both parapsychologists and their critics, has never been found in parapsychological investigations in the history of psychic research. Recently a few parapsychologists have expressed the hope that the experiments on remote viewing, random number generators, and the Ganzfeld (the very ones we have chosen to examine in detail in this report) may actually yield the long-sought replicability. The type of replicability that has been claimed so far is the possibility of obtaining significant departures from the chance baseline in only a proportion of

the experiments, which is a kind of replicability quite different from the consistent and lawful patterns of covariation found in other areas of inquiry.

Despite the fact that scientific progress in a given area depends on the accumulation of lawful and consistent patterns across many experiments, the methods for deciding that such consistency exists are still quite primitive in comparison with the standards for judging the adequacy of a single experiment. Indeed, it is only within the past few years that serious attention has been devoted to developing objective and standardized procedures for evaluating the consistencies across a body of independent studies. For the most part, judgment about what a body of investigations demonstrates is still a surprisingly intuitive and haphazard process. This probably has not been a serious drawback in those areas of inquiry in which the basic phenomena are robust and experiments can be conducted with high confidence that the predicted relations will be obtained; but such impressionistic means for aggregating the outcomes of several experiments in the domain of parapsychology open the door to all the motivational and cognitive biases discussed in the paper prepared for the committee by Griffin. Not only are the data and alleged correlations erratic and elusive in this field, but their very existence is open to question.

EVALUATION OF THE SCIENTIFIC EVIDENCE

To evaluate the best scientific evidence on the existence of psi, and with the advice of proponents and our sponsors, we conducted site visits to some of the most notable parapsychological laboratories. The parapsychology subcommittee (see Appendix C) visited Robert Jahn's Engineering Anomalies Research Laboratory at Princeton University, where it witnessed presentations and demonstrations regarding psychokinetic experiments on random number generators. Jahn and his associates also briefed the subcommittee on the current status of their work in remote viewing.

The subcommittee also visited Helmut Schmidt's laboratory at the Mind Science Foundation, San Antonio, Texas. Schmidt pioneered the use of random number generators in parapsychology experiments in 1969. His is considered one of the two major research programs on psychokinesis (the second is Jahn's).

As an additional posssible input, the committee agreed to participate in a psychokinetic experiment of new design with Helmut Schmidt. Specifically, Schmidt accepted the suggestion that the committee's consultant, Paul Horwitz, be included in the conduct of the experiment. The

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

work has not yet begun, however, and it now appears that we will not have any results to report before our terms expire.

The chair of the parapsychology subcommittee also visited SRI International, another major laboratory studying psychic effects on random number generators. (This latter research group argues that the observed effects are not due to psychokinesis but rather represent a special form of precognition.) The subcommittee chair also attended the meetings of the Parapsychological Association held at Sonoma State College in California. The entire committee made a site visit to Cleve Backster's laboratory in San Diego (arranged to coincide with the committee's meeting in La Jolla, California).

These site visits enabled the committee to observe firsthand the experimental arrangements and equipment used by some of the major contributors to parapsychological research. They also provided us an opportunity to discuss results, interpretations, and problems with a few important investigators. We were impressed with the sincerity and dedication of these investigators and believe that they are trying to conduct their research in the best scientific tradition. We also got the impression that this type of research involves many unresolved problems and still has a long way to go before it develops standardized, easily replicable procedures. The information obtained from these site visits does not provide an adequate basis for making scientific judgments. For this we rely, as we would in other fields of science, on a careful survey of the literature.

Research on Remote Viewing

The SRI Remote Viewing Program

Since the early 1970s, probably the best known research program in parapsychology has been the experiments in remote viewing initiated by physicists Harold Puthoff and Russell Targ when they were at SRI International. In a typical remote viewing experiment a subject, or percipient, remains in a room or laboratory with an experimenter, while a target team visits a randomly selected geographical site (e.g., a shopping mall, an outdoor arena, the Palo Alto airport, the Hoover tower). Neither the experimenter nor the subject has been given any information about the target. Once the experimenter and the subject are closeted in the laboratory, they wait for 30 minutes before the subject begins to describe his or her impressions of the target site.

Meanwhile the target team, consisting of two to four members of the SRI staff, obtains instructions for going to a randomly chosen target site from another SRI staff member. They then drive to the designated target site and remain there for an agreed-on 15-minute period (after allowing approximately 30 minutes to reach the site). During the time that the target team remains at the target site, the subject describes his or her impressions into a tape recorder and also makes any drawings that would help to clarify those impressions. When the target team returns to the laboratory, all the participants listen to the tape recording of the subject's impressions. Then all the participants go to the target site, where the subject is allowed to see how closely his or her impressions agreed with the actual target.

The first subject to participate in such a formal series of trials was the late Pat Price. In the first series, consisting of nine sessions, the duration of each session was 30 minutes. The transcript for each session is rich in detail; the one published transcript in Targ and Puthoff's first book runs to almost six printed pages (Targ and Puthoff, 1977).

Given such data, how does one decide if the experiment was a success? Did Price's descriptions, for example, convey correct knowledge of the different target sites? In fact, two methods have been used to demonstrate the effectiveness of remote viewing. One method is simply to compare the description with the target and make a judgment as to whether the correspondence is sufficient to claim a "hit." The second method uses an independent judge to rank the degree to which each description matches each site and then applies statistical tests to decide if the association is greater than chance.

Unprecedented success was claimed for the early remote viewing experiments in terms of both methods (Targ and Puthoff, 1974, 1977; Puthoff and Targ, 1976). Many examples were supplied of dramatic correspondences between impressions of the percipient and the physical details of the actual target. Such correspondences, no matter how dramatic and compelling, do not carry scientific weight, because it is impossible to assess their probabilities. In addition, much psychological research indicates how such subjective validation can create strong, but false, illusions of matching (see below).

The more formal evidence from the rankings of independent judges was also impressive. The first formal series of nine trials resulted in seven of the transcripts being ranked 1 against their intended target sites by the independent judge. Only one such ranking would be expected by chance. Puthoff and Targ reported the probability of such an outcome being due to chance as only 0.0000029. The second formal series, using Hella Hammid, was equally impressive, producing five first places and four second places in the rankings of transcripts against target sites.

Although subsequent series by Targ and Puthoff, as well as by

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

other investigators, have not always yielded such overwhelmingly impressive results, most of them have continued to display highly significant outcomes (Targ and Harary, 1984). On the surface, at least, this is a reliable, simple, and highly effective recipe for producing paranormal communication. Especially appealing is the claim that remote viewing works with just about everyone. Targ and Harary, for example, provide exercises for anyone who wants to develop and improve his or her ability to pick up information at remote sites. Neither space nor time, its proponents assert, is a barrier. The percipient can pick up information from the surface of Jupiter as well as from target sites that can be visited at some future time.

Scientific Assessment of Remote Viewing

After the first remote viewing experiments were conducted in the early 1970s, many investigators throughout the world tried to follow suit. Most of them believed that their findings supported the claims of the SRI International researchers. The majority of these experiments, however, consisted of informal demonstrations rather than formal scientific experiments and relied solely on subjective matching. In the past 15 years, the number of formal experimental replications of the SRI remote viewing experiments has been surprisingly few.

Targ and Harary (1984) include as an appendix in their book a report by Hansen, Schlitz, and Tart that evaluates all the known remote viewing experiments conducted from 1973 through 1982. "In an examination of the twenty-eight formal published reports of attempted replications of remote viewing," write Targ and Harary, "Hansen, Schlitz, and Tart at the Institute for Parapsychology found that more than half of the papers reported successful outcomes." They concluded: "We have found that more than half (fifteen out of twenty-eight) of the published formal experiments have been successful, where only one in twenty would be expected by chance."

Two comments may be in order with respect to the foregoing conclusion. First, given the enormous publicity and the unusually strong claims, 28 formal experiments in 10 years seems surprisingly few. In comparison, the Ganzfeld psi experiments produced approximately twice as many formal experiments during the same interval. Second, 13 of the 28 formal experiments, or 46 percent, failed to claim successful outcomes. This rate of failure is much higher than what might have been expected on the basis of the earlier claims by Targ and Puthoff (1977), namely, that they had succeeded with every subject they had tried.

Even 15 successful outcomes out of 28 tries is impressive, especially by parapsychological standards. An inspection of the listed studies, however, suggests that the 28 formal experiments vary considerably in their importance. Some of these "published formal experiments" appeared as brief reports or abstracts of papers delivered at meetings of the Parapsychological Association or similar organizations. Others appeared in print only as brief or informal reports in book chapters or letters to the editor. Altogether, 15 of the 28 were published under conditions that fall short of scientific acceptability. Only 13, or 46 percent, of the experiments were published under refereed auspices. As in other sciences, only published reports that have undergone peer review and are adequately documented can be considered seriously as part of the scientific data base.

Of the 13 scientifically reported experiments, 9 are classified as successful in their outcomes by Hansen et al. (Targ and Harary, 1984). Seven of these nine experiments were conducted by Targ and Puthoff at SRI International, the remaining two at other laboratories. This relatively small harvest of nine "successful" experiments suffers from the fact that each is seriously flawed. A variety of problems afflicts the published reports on remote viewing. The documentation, even according to many parapsychologists, is seriously inadequate. Attempts by both neutral and skeptical investigators to gain access to the raw data have typically been thwarted or strongly resisted. Because the essence of scientific justification is public accessibility to the data, this relative inaccessibility suggests that much of the remote viewing data base is not part of science.

Most of the reasons for questioning the acceptability of the evidence for remote viewing lie in a methodological flaw that characterizes all but one of the experiments deemed successful: the successive trials are not independent of one another. This lack of independence has unfortunate consequences for any attempt to draw conclusions about ESP based on the outcomes of such experiments. The concept of independence is technical and somewhat difficult to explain simply, but, since it is critical to understanding why the remote viewing experiments fail to make their case, we supply an intuitive explanation.

Assume that we are considering a remote viewing experiment in which the subject participates in only two trials. In other words, we deal with two randomly chosen target sites. For the first trial, the target team goes to the first target site and remains there while the subject produces his or her first description. Immediately after this trial, the target team returns to the laboratory and takes the subject to the actual target site so that he or she and the others can gain a

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

subjective impression of how closely the description corresponds with the target. For the second trial, the target team visits a second randomly chosen site. While they are visiting this site, the subject produces a second description.

When the experiment is over, the list of target sites (in random order) and the transcripts of the subject's descriptions are given to a judge, who also visits each site. While at a given site, the judge reads the two transcripts and ranks them in terms of how well each one corresponds with the particular site. In our example, one of the transcripts will be ranked 1 and the other will be ranked 2 (with 1 indicating the better correspondence between that target and the transcript). After visiting one site and doing this ranking, the judge then visits the second site and repeats the ranking procedure. The raw data can be set out in a matrix with the target sites as the columns and the transcripts as the rows.

A perfect outcome would be indicated if the transcript produced at the time the team was visiting site A was ranked 1 against that site, and the transcript produced when the team was visiting site B was ranked 1 for that site. (Of course, two trials would be too few to make an adequate statistical assessment of the success of the matching—successful matching would occur too frequently just by chance. The principles we want to illustrate, however, remain the same for two as for many trials.)

If the successive trials in the experiment were independent of one another, and we were interested only in direct hits (that is, outcomes for which the intended transcript was rated 1 against the target site), then we could expect the subject to make between zero and two direct hits. Indeed, if chance alone were operating, there would be four, equally likely, possibilities: (1) no hits, (2) a hit on the first trial and a miss on the second, (3) a miss on the first trial and a hit on the second, and (4) two hits. By this reckoning, the subject could be expected to get two direct hits just by chance in one of every four experiments.

But, as we indicated, the successive trials are not independent. This is because the judge is almost certainly not going to rank a transcript as 1 for more than one target site. This means, in our example, that if he or she ranks the first transcript 1 for target A, then he or she will probably rank the second transcript 1 for target B. In effect, this lack of independence between trials means that, instead of four equally likely possible outcomes there are only two: no hits or two hits. The dependence between trials has created a situation in which the chance probability of two hits is now 50 percent rather than 25 percent.

In this situation, if an experimenter uses a statistical test that assumes independence, he or she will come out with the wrong probabilities. In fact, the statistical test will exaggerate the significance of many outcomes. The failure of the experimenters to realize this problem resulted in exaggerated levels of significance for the early remote viewing experiments. Kennedy (1979), who originally pointed to this problem, recalculated the probabilities for some of these experiments. Puthoff and Targ (1976) reported that five of their first six remote viewing experiments were significant at the .05 level. With Kennedy's corrections for lack of independence, only two remained significant. According to Kennedy, only one of the two successful replications by Bisaha and Dunne (1979) remained significant with the more appropriate test.

One reason for the optimistic initial beliefs in the scientific reality of remote viewing was the fact that the lack of independence between trials produced exaggerated odds against chance results. But even with conservative corrections for lack of independence, approximately one-third of the early experiments still yielded successful outcomes.

One easy way to avoid this problem of dependence is to use a separate target pool of possible sites for each trial. For example, for the first trial one could designate a pool of four possible sites, one of which is randomly chosen to be the actual target site. A second pool of four different possible sites would be used for the second trial. When the trials are completed, the judge is given the list of the four sites for the first trial along with the subject's description for that trial. The judge then ranks each site in terms of its correspondence to the description. The four possible sites for the second trial are then ranked in terms of their correspondence to the subject's description for the second trial. In this illustration, the subject has a probability of 1 in 4 of having the actual target site ranked 1 on each trial, or a probability of 1 in 16 of being correct on both trials.

This second procedure, which is typically used in most free-response parapsychological experiments (such as the Ganzfeld experiments discussed below), not only guarantees independence between successive trials, but also avoids other serious problems, which we discuss next. The fact that the subject is given feedback by being taken to the target site immediately after each trial creates an additional form of dependence between trials. For this reason, other possibilities exist for obtaining "successful" results artifactually. The transcripts can contain clues that provide nonparanormal reasons for judges to associate descriptions with targets correctly. Some of these

clues can be quite overt, such as when a subject mentions in the description how the current target apparently differs from a previous target site. When such a clue appears in the description, it provides the judge with information that the current description does not belong with the previous site. This increases the probability that the description will be matched with its appropriate target.

Marks and Kammann (1978) initiated a controversy, still not fully resolved, by claiming that such overt clues were sufficient to account for the striking results of the very first SRI remote viewing with Pat Price. Targ and Puthoff did not deny the existence of such clues in the Price series but argued that they were not sufficient to have accounted for the results. This dispute still has not been settled (Targ, Puthoff, and Targ, 1980; Scott, 1982; Marks and Scott, 1986).

Possibly this controversy over the role of the more overt clues has deflected attention from a much more fundamental and fatally damaging criticism first made by Hyman (1979) and independently by Kennedy (1979). Hyman and Kennedy pointed out that the combination of immediate feedback and lack of independence between successive trials makes it virtually impossible to prevent sensory cueing in the transcripts. As long as both the subject and the experimenter who is closeted with the subject are not blind to the preceding target sites, there is no way to prevent the transcript from being affected in a variety of possible and perhaps subtle ways by the knowledge of the preceding targets.

Hyman (1984–1985) provides an illustration of how such implicit sensory cueing might occur (pp. 131–132):

> Say that the target for the first session was the Hoover Tower at Stanford. This will almost certainly influence what both the viewer and the interviewer say during the second and subsequent sessions in the same series. Almost certainly the viewer, during the second session, will not supply an exact description of the Hoover Tower. So, whatever the viewer says during the second session, a judge should find it to be a closer match to the second target site than to the first one. Now, assume that the second target site happened to be the Palo Alto train station. The viewer's descriptions during the third session will avoid describing either the Hoover Tower or the Palo Alto train station. We do not need to hypothesize something as mysterious as psi to predict that a judge should find this third description a better match to the third target site than either of the first two. As we add sessions, this effect of immediate feedback should continue to make the correlation between the viewer's descriptions and the target sites better and better.

No amount of editing for overt clues can overcome this defect of remote viewing experiments that follow the SRI pattern of dependent trials and immediate feedback. The mechanism described by Hyman

should result in some dramatic correspondences. These dramatic correspondences, in conjunction with subjective validation, are a highly potent recipe for creating the illusion (for both experimenters and subjects) that ESP has occurred.

Palmer (1985), a major parapsychologist who otherwise carefully considers the criticisms of parapsychology, misses the seriousness of this flaw. In mentioning Hyman's criticism, he writes (p. 50):

> It has been suggested by Hyman (1979) that since the subjects in most cases received feedback of the correct target after each trial, the subject could have gained some advantage by avoiding to mention characteristics of targets in earlier trials in their responses in later trials. As noted by Targ, Puthoff, and May (1979), the target pool for the geographical-site experiments was sufficiently large and contained sufficient redundancy that this is unlikely to be a significant biasing factor.

Perhaps such complacency has enabled experimenters to continue conducting remote viewing experiments with this fatal flaw. In fact, the size of the target pool, no matter how large, does not affect the validity of Hyman and Kennedy's criticism. Nor does the claim that the pool contained sufficient redundancy make much difference. Each geographical site is unique and contains a combination of specific characteristics that distinguishes it from the other sites in a given series. Indeed, as the parapsychologists themselves have asserted, unless this were so, there would be no possibility of the transcripts' being uniquely associated with a given target site. In every one of the remote viewing experiments that allows the possibility of subtle cueing, the possibility of the judges' being able to make completely successful matchings because of this artifact is highly plausible; and as long as a highly plausible, normal alternative to ESP can account for the apparent success of the outcomes the parapsychologists, by their own standards, cannot claim evidence for paranormal transmission of information.

As it turns out, all but one of the nine scientifically reported studies of remote viewing (at the time of the Targ and Harary survey) suffer from the flaw of sensory cueing. The one experiment that cannot be faulted for this reason is the long-distance remote viewing experiment of Schlitz and Gruber (1980). However, as Hyman (1984–1985) has pointed out, this experiment suffers from another very serious flaw. Gruber, who was a member of the target team and thus was familiar with the targets, translated the subject's target descriptions into Italian for the judging process. Why the experimenters allowed such potential sources of biased experimental procedures is not known, but the violation obviously negates the results as evidence for psi.

Since the Targ and Harary survey, we have learned of two attempts

to replicate the Schlitz and Gruber experiment without the flaw mentioned. One, still unpublished, produced negative results. The second, by Schlitz and Haight (1984), produced marginally significant results. Indeed, if the more acceptable two-tailed test of significance had been used, the results would not have been considered significant by customary standards. Although the report of this study lacks sufficient documentation with respect to certain aspects of procedure, both Palmer (1985) and Alcock agree that this is the best controlled and most methodologically sound of all the remote viewing experiments so far.

In summary, after approximately 15 years of claims and sometimes bitter controversy, the literature on remote viewing has managed to produce only one possibly successful experiment that is not seriously flawed in its methodology—and that one experiment provides only marginal evidence for the existence of ESP. By both scientific and parapsychological standards, then, the case for remote viewing is not just very weak, but virtually nonexistent. It seems that the preeminent position that remote viewing occupies in the minds of many proponents results from the highly exaggerated claims made for the early experiments, as well as the subjectively compelling, but illusory, correspondences that experimenters and participants find between components of the descriptions and the target sites.

Research on Random Number Generators

The Basic Paradigm

The use of random number (or random event) generators for parapsychological research began in the 1960s and became relatively standard during the 1970s as the technology became widely available. A random number generator (RNG) is simply an electronic device that uses either radioactive decay or electronic noise to generate a sequence of random symbols. Originally such devices were used to test ESP, usually clairvoyance or precognition, but the most widespread and widely known work focuses on what is called micropsychokinesis, or micro-PK. In such research a subject, or operator, attempts to mentally bias the output of the random number generator, so that it produces a nonrandom sequence.

Most of the work with RNGs has used binary generators, or what Schmidt calls "electronic coin flippers." The output on each trial is either 0 or 1, that is, heads or tails. If the RNG is unbiased and truly random, then it should produce, on control runs, sequences of 0s and 1s that are independent of each other and that, in the long run, will yield 1s 50 percent of the time.

In a typical experiment, a subject (either a person who claims to be a psychic or a person chosen for availability who does not make such claims) is placed in the vicinity of the RNG and attempts to bias the output either toward more or fewer 1s. When an animal is used as the subject, the RNG output is usually coupled to an outcome whose frequency the animal presumably would like to either increase or decrease. In an experiment carried out with cockroaches, for example, one outcome was electric shock. If, during the time the output of the RNG was coupled with the shock apparatus, the proportion of shocks decreased below 50 percent, this would be taken as evidence of a psychokinetic effect of the cockroach on the output of the RNG.

The RNG experiments have been of interest to some military and governmental personnel because of the possibility, if such micro-PK is demonstrable, of psychically affecting equipment and computers that depend on the output of electronic symbols.

Results of the Experiments

In a recent survey 56 reports published between 1969 and 1984 and dealing with research on possible psychokinetic perturbations of binary RNGs (Radin, May, and Thomson, 1985), the reviewers counted 332 separate experiments. Of the 332 experiments, 188 were reported in refereed journals or conference proceedings, and of these 188 experiments with some claim to scientific status, 58 reported statistically significant results (compared with the 9 or 10 experiments that would be expected by chance). The other 144 experiments were produced by the Engineering Anomalies Research Laboratory at Princeton University; none of them had been published in a refereed journal at the time of the survey. Of these 144 experiments, 13 were classified as yielding statistically significant results. So, in the total sample of 332 experiments, 71 yielded ostensibly significant results at the traditional .05 level. This amounts to a success rate of approximately 21 percent, compared with the rate of 5 percent that would be expected by chance.

Palmer (1985) and Alcock agree that such results cannot be accounted for by chance. In other words, both the parapsychologist and the skeptic, in their respective reviews of the RNG research, agree that something other than accidental fluctuation is producing these results. Palmer calls this something an anomaly, which, while it may or may not be paranormal, cannot be explained by current scientific theories. Alcock points to various defects in the experimental protocols and concludes that no conclusions about the origins of these departures from randomness are justified until successful

outcomes can be more or less consistently produced with adequately designed and executed experiments.

Both Palmer and Alcock focus their reviews on the two most influential research programs on RNGs. One is the program of Helmut Schmidt, a quantum physicist who began working on psi and RNGs in 1969. The other is the program begun by Robert Jahn in the late 1970s, when he was dean of the School of Engineering and Applied Science at Princeton University (see Jahn, 1982). These two programs have accounted for almost 60 percent of all known experiments on RNGs. They have also been the most consistently successful in achieving statistically significant outcomes.

Although the results suggest that on each experimental group of trials the number of 1s is greater or less than the 50 percent baseline (depending on the intended direction), the actual degree of deviation from chance is quite small. As Palmer (1985) indicates, Schmidt's subjects have averaged approximately 50.5 percent hits over the years, compared with the expected baseline of 50 percent. This amounts to producing one extra 1 every 100 trials. The reason such a small departure from chance is statistically significant is that an enormous number of trials is conducted with each subject.

Jahn and his colleagues at Princeton have, in a much shorter time, produced on the order of 200 times the number of trials that Schmidt did in 17 years. The Princeton researchers have also produced a significantly lower success rate than Schmidt. In their formal series of 78 million trials, the percentage of hits in the intended direction was only 50.02 percent, or an average of 2 extra hits every 2,500 trials. Again, such an extremely weak effect is statistically significant only when one is dealing with very large numbers of trials.

Scientific Assessment of the RNG Experiments

Palmer (1985) carefully reviews the major criticisms of the work of Schmidt and Jahn. He addresses questions about security, because subjects often are left alone with the apparatus during the data collection. In the Princeton experiments, the data are always collected when the subject is alone with the apparatus. Although the Princeton experiments now contain a number of features that would make it extremely difficult for a naive subject to bias the results, it is not clear that this has always been so. It would make good scientific sense to conduct some trials during which the subject is carefully monitored to see if successful outcomes are still obtained.

The major reservations about the RNG experiments concern the adequacy of the randomization of the outputs. Schmidt applied only limited tests for the randomness of his machines, and most of the

control trials were gathered by allowing the machine to run for long periods, usually overnight. Although these controls usually produced results in line with the chance baseline, critics have pointed out that the controls are unsatisfactory because they were not conducted for shorter runs and at the same time as the data from the experimental sessions.

Palmer grants that the critics are correct in pointing out some of the shortcomings in Schmidt's methods for testing and controlling for the randomization of his machines. Palmer also correctly points out that such criticism is somewhat blunted by the fact that the critics have not specified any plausible mechanisms that would account for the obtained differences between the experimental and control trials. He is correct in pointing out that the Princeton experiments provide more adequate controls; however, he has probably assumed that the baseline controls in the Princeton experiments were run at the same time as the two experimental conditions of hitting and missing. It is easy to interpret the somewhat ambiguous description of the procedure in this manner. The relevant part of the authors' methodological description is as follows (Nelson, Dunne, and Jahn, 1984:9):

> The primary variable in these experiments is the operator's pre-recorded intention to shift the trial counts to higher or lower numbers. This directional intention may be the operator's choice—the so-called "volitional" mode—or it may be assigned by a specified random process—the "instructed" mode. In either mode, data are collected in a "tri-polar" protocol, wherein trials taken under an intention to achieve high numbers (PK+), trials taken under an intention to achieve low numbers (PK−), and trials taken as baseline, i.e. under null intention (BL), are interspersed in some reasonable fashion, with all other operating conditions held identical. For all three streams of data, effect size is measured relative to the theoretical chance mean. This tri-polar protocol is the ultimate safeguard in precluding any artifacts such as residual electronic biases or transient environmental influences from systematically distorting the data.

At first glance it might appear as if the tripolar protocol requires that the two types of experimental groups of trials and the baseline group of trials always be taken at the same session. This would be consistent with the claim that "any artifacts such as residual electronic biases or transient environmental influences" were thereby precluded "from systematically distorting the data." Such a claim would be justified if, in fact, at each session one group of trials of each of the three types was obtained, provided that each group of trials was of the same length and that the order of the three types of trials was independently randomized for each session.

The description provided by Nelson and his colleagues says nothing

at all about the order in which the three conditions were conducted, and a careful reading indicates that the baseline data may not always have been obtained at the same sessions and under the same conditions as the experimental groups of trials. It is not clear what the authors mean by stating that the three trials "are interspersed in some reasonable fashion." In fact, an examination of the data reported for each subject makes it clear that the strict tripolar protocol could not possibly have been followed with much of the data collection, because in many cases the baseline data are entirely absent or occur with many fewer trials than the experimental data. Indeed, it is not even clear that PK+ and PK− trials were always obtained at the same sessions, because for some subjects the total numbers of these trials are not equal.

We suspect that, over the six years or so during which the Princeton group was accumulating its data base, it made many changes in both the hardware and the experimental protocol. The sophisticated procedures currently in use and the requirement that the three types of trials be of equal length and that one of each be conducted at each session are the most recent variations in the paradigm. Unfortunately, the data are not presented in such a way that it is possible to determine whether the successful results are due to the earlier or the later experiments.

Such issues become especially important when we consider the extremely small size of the effect being claimed and when we further realize, as Palmer has pointed out, that the bulk of the significance in the formal series was due to just one subject, who contributed 23 percent of the total data. This one subject achieved a hit rate of 50.05 percent. When her data are eliminated, the remaining data yield a hit rate of 50.01 percent, which is no longer significantly different from chance.

In other words, it looks as if almost all the success of Jahn's huge data base can be attributed to the results from one individual, who, over the years, produced almost 25 percent of the data. This one individual was not only the most experienced subject, but also, presumably, familiar with the equipment. When combined with the fact, as Palmer points out, that the Princeton experiments provide inadequate documentation on precautions to prevent tampering by subjects, it becomes even more important to see if the same degree of success can be achieved when the sessions are adequately monitored.

Alcock, in his review of the same RNG studies surveyed by Palmer, points to a number of weaknesses in both the Schmidt and the Princeton experiments. For example, he faults Schmidt's experiments for such things as inadequate controls, failure to examine the target sequences, overcomplicated experimental setups, inadequate tests of randomness, and lack of methodological rigor. Alcock faults the Princeton experiments for such things as failing to randomize the sequence of groups of trials at each session, inadequate documentation on precautions against data tampering, and possibilities of data selection.

Palmer and Alcock do not really differ in their assessments of the shortcomings of the Schmidt and Princeton RNG experiments. They do differ, however, on what conclusions can be drawn from such imperfect experiments. Palmer emphasizes the fact that the critics have not provided plausible explanations as to how the admitted flaws could have caused the observed results. His position seems to be that, unless the critics can provide such plausible alternatives, the results should be accepted as demonstrating an anomaly. Alcock focuses on the fact that the successful results have been obtained under conditions that fall short of the experimental ideals that parapsychologists themselves profess. He emphasizes that the parapsychologists have no right to claim to have demonstrated psi from experiments that have been conducted with "dirty test tubes." Such a revolutionary conclusion as the existence of psi demands justification from experiments that have clearly used "clean test tubes."

What would it take to conduct an adequate RNG experiment? May, Humphrey, and Hubbard (1980) set out to do just that. After reviewing all available RNG experiments from 1970 through 1979 and taking into account the various deficiencies in these experiments, they gathered together and meticulously tested the components necessary to provide adequately randomized trials. They also devised a careful experimental protocol and set out in advance the precise criteria that would have to be fulfilled before they could call their results successful. Going further, after they completed the experiment with results that met their criteria for success, they subjected their equipment to all sorts of physical extremes to see if they could obtain such a degree of success by a possible artifact.

They report that this singularly well controlled RNG experiment in fact met their criteria for success. It is unfortunate, therefore, that this carefully thought-out experiment was conducted only once. After the one successful series, using seven subjects, the equipment was dismantled, and the authors have no intention of trying to replicate it (personal communication, August 1986). It is unfortunate because this appears to be the only near-flawless RNG experiment known to us, and the results were just barely significant. Only two of the seven subjects produced significant results, and the test of overall significance for the total formal series yielded a probability of 0.029.

The experiment, while nearly flawless, still had some problems as evidence for psi. For one thing, it was reported only in a technical report in 1980 and has never been published in a refereed scientific journal. Despite the admirable attention to details, all the control trials were taken when no human being was present. One might argue that this was not an ideal control for the experimental session, in which a subject was physically present in the room. The authors have assured us that their various attempts to bias the machine by physical means almost certainly rule out the possibility that the mere presence of a human being could have affected the output. However, a physicist who claims to have several years of experience in constructing and testing random number devices tells us that it is quite possible, under some circumstances, for the human body to act as an antenna and, as a result, possibly bias the output.

May and his colleagues at SRI, in the same technical report in which they claim successful results for their single experiment, surveyed all the RNG experiments known to them through the year 1979 and found that their combined significance was astronomically high. They add (May, Humphrey, and Hubbard, 1980:8):

> This impressive statistic must, however, be evaluated with respect to experimental equipment and protocols. All the studies surveyed could be considered incomplete in at least one of the following four areas: (1) No control tests were reported in more than 44 percent of the references. Of those that did, most did not check for temporal stability of the random sources during the course of the experiment. (2) There were insufficient details about the physics and constructed parameters of the experimental apparatus to assess the possibility of environmental influences. (3) The raw data was not saved for later and independent analysis in virtually any of the experiments. (4) None of the experiments reported controlled and limited access to the experimental apparatus.

As far as we can tell, the same four points can be made with respect to the RNG experiments that have been conducted since 1980. The situation for the RNG experiments thus seems to be the same as that for remote viewing: over a period of approximately 15 years of research, only one successful experiment can be found that appears to meet most of the minimal criteria of scientific acceptability, and that one successful experiment yielded results that are just marginally significant.

Research on the Ganzfeld

The Ganzfeld Experiments

The Ganzfeld psi experiments are named after the term used by Gestalt psychologists to designate the entire visual field. For theoretical purposes, the Gestalt psychologists wanted to create a situation in which the subject or observer could view a homogeneous visual field, one with no imperfections or boundaries. Psychologists later discovered that when individuals are put into a Ganzfeld situation they tend quickly to experience what they described as an altered state of mind.

In the early 1970s, some parapsychologists decided that the use of the Ganzfeld would provide a relatively safe and easy way to create an altered state in their experimental subjects. They believed that such a state was more conducive to picking up the elusive psi signals. In a typical psi Ganzfeld experiment, the subject, or percipient, has halved ping-pong balls taped over the eyes. The subject then reclines in a comfortable chair while white noise plays through earphones attached to his or her head. A bright light shines in front of the subject's face. When seen through the translucent ping-pong balls, the light is experienced as a homogeneous, foglike field. When so prepared, almost all subjects report experiencing a pleasant, altered state within 15 minutes.

While one experimenter is preparing the subject for the Ganzfeld state, a second experimenter randomly selects a target pool from a large set. The target pool typically consists of four possible targets, usually reproductions of paintings or pictures of travel scenes. One of the four is chosen at random to be the target for that trial. The target is given to an agent, or sender, who tries to communicate its substance psychically to the subject in the Ganzfeld state. After a designated period, the subject is removed from the Ganzfeld state and presented with the four candidates from the target pool. The subject then ranks the four candidates in terms of how well each matched the experience of the Ganzfeld period. If the actual target is ranked first, the trial is designated a hit. An actual experiment consists of several trials. In the example, the probability is that one of every four trials will produce a hit. If the number of hits significantly exceeds the expected 25 percent, then the result is considered to be evidence for the existence of psi.

Critique of the Ganzfeld Experiments

In a careful and systematic review of the Ganzfeld experiments undertaken in 1981 and published in the March 1985 issue of the *Journal of Parapsychology*, Hyman concluded that the data base exhibited flaws involving multiple testing, inadequate controls for sensory leakage, inadequate randomization, statistical errors, and inadequate documentation. These flaws, in his opinion, were sufficient

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

to disqualify the Ganzfeld data base as evidence for psi. Of the 42 experiments, 39 (93 percent) used multiple analyses, which artificially inflated the chances of obtaining significant outcomes. Only 11 (26 percent) clearly indicated that they had adequately randomized the target selections. As many as 15 (36 percent) used inferior randomization, such as hand shuffling, or no randomization at all. The remaining 16 experiments did not supply sufficient information on how they had chosen the targets. As many as 23 of the experiments (55 percent) used only one target pool, which means that the subject was handed for judging not a copy of the target but the very same target that the percipient had handled, permitting the possibility of sensory cueing. Although the argument for psi is mainly a statistical one, the reports of 12 experiments (29 percent) revealed statistical errors. A number of other departures from optimal practice were also found.

The same issue of the *Journal of Parapsychology* contained a lengthy rebuttal by parapsychologist Charles Honorton, one of the pioneers of the Ganzfeld psi technique. Honorton disputed many of Hyman's opinions as to what constituted flaws; provided a reanalysis of the data base to overcome many of the statistical weaknesses of the original experiments; and argued that the flaws he agreed existed were not sufficient to have accounted for the findings. In this respect his analysis is consistent with Palmer's approach. He does not deny that the experiments depart from optimal design, but he argues that such departures are insufficient to account for the results.

Honorton and Hyman had the opportunity to discuss their differences about psi in general at the Parapsychological Association meetings in 1986; as a result, they agreed to draft a joint communiqué to emphasize those points on which they agree. That communiqué appeared in the December issue of the *Journal of Parapsychology* (Hyman and Honorton, 1986). They agree that the current data base is insufficient to support either the conclusion that psi exists or the conclusion that the results are due to artifacts. They further agree that the issue can be settled only by future experiments conducted according to the stated standards of parapsychology, which are also the accepted standards of psychological research.

Another important input to the committee's judgment on the Ganzfeld research was the systematic evaluation of the contemporary parapsychological literature by Charles Akers (1984), a former parapsychologist. Akers's critique used a methodological strategy different from that used by Hyman. Hyman undertook to evaluate the entire data base of a single research paradigm (Ganzfeld), including both successful and unsuccessful outcomes. Akers surveyed

contemporary ESP experiments broadly, but confined his evaluation to those that had produced significant results with unselected subjects. Hyman assigned flaws to experiments without regard to whether each flaw, by itself, could have caused the observed outcome. Akers charged a flaw to a study only if he thought the flaw could have been sufficient to produce the observed result. He chose a sample of 54 parapsychological experiments from areas of research that had been previously reviewed by Honorton or Palmer; his intent was to choose experiments that could be viewed as the best current evidence for the existence of psi. As a result of this exercise, he concluded (Akers, 1984:160–161):

> Results from the 54-experiment survey have demonstrated that there are many alternative explanations for ESP phenomena; the choice is not simply between psi and experimenter fraud. . . . The numbers of experiments . . . flawed on various grounds were as follows: randomization failures (13), sensory leakage (22), subject cheating (12), recording errors (10), classification or scoring errors (9), statistical errors (12), reporting failures (10). . . . All told, 85% of the experiments were considered flawed (46/54).
>
> This leaves eight experiments where no flaws were assigned. . . . Although none of these experiments has a glaring weakness, this does not mean that they are especially strong in either their methods or their results. . . .
>
> In conclusion, eight experiments were conducted with reasonable care, but none of these could be considered as methodologically ideal. When all 54 experiments are considered, it can be stated that the research methods are too weak to establish the existence of a paranormal phenomenon.

Research on Electrical Activity and Emotional States

The Backster Laboratory

In addition to examining parapsychological research in areas that have produced large literatures, the committee witnessed an example of experimental work at a far less developed stage. On February 10, 1986, committee members visited the Backster Research Foundation in San Diego and saw a demonstration of experimental procedures for detecting a correlation between the electrical activity of oral leukocytes and the emotional states of the donor.

Cleve Backster is a polygraph specialist who had at one time helped develop interrogation techniques for the Central Intelligence Agency and now runs his own polygraph school in San Diego. The school is housed in the same rooms that constitute the Backster Research Foundation, which is devoted to the study of what Backster refers to as primary perception. Backster's research on paranormal matters

began in February 1966, when he recorded, from a philodendron plant that he had hooked up to a polygraph, a response he recognized as similar to that of human beings in emotional states. Backster believed he had demonstrated that the plant showed such emotional response when brine shrimp or other living organisms were either threatened or actually killed in an adjoining room. The notion of primary perception in plants became both a popular subject for research and a highly controversial concept during the late 1960s and early 1970s.

We were told that Backster has quietly continued his researches into this and related matters. He has now devised a technique for recording electrical activity in leukocytes taken from a donor's mouth. The advantage of this technique, we were told, is that the leukocytes respond mostly to emotional states of the donor.

One committee member volunteered to be the demonstration subject. Another member accompanied him to observe the techniques for obtaining the leukocytes and preparing them for recording. The sample was obtained by having the subject "chew" on a 1.2 percent saline solution and then spit it back into a centrifuge tube. Ten such samples were obtained in this way. The samples were then spun in a centrifuge for six minutes, and the particulate matter at the bottom of each tube was pipetted into the preparation tube. The preparation tube contained about one centimeter of particulate matter and was filled almost to the top with 1.2 percent saline solution. Two uninsulated wire electrodes were inserted into the bottom of the tube, which was then placed within a shielded cage and connected by leads to an EEG-type recording apparatus.

During the demonstration, the subject sat approximately two meters from the preparation. We were told that subjects usually sit about five meters from the preparation. A split-screen projection video display was provided: the lower portion of the screen recorded the movements of the polygraph paper and pen as they produced a record of the electrical activity presumably taking place in the leukocyte preparation. The upper portion of the screen recorded the behavior of the seated subject.

In his previous research using this arrangement, Backster reported that, when the subject revealed an emotional reaction, the electrical action of the leukocytes showed a corresponding reaction. During our demonstration, the polygraph record produced several strong deflections in both the control and the experimental series, but they did not obviously correlate with any corresponding thoughts or emotional states of the subject as various stimuli were presented. Backster suggested that this was probably because so many people were crowded into the laboratory that the leukocytes were responding to thoughts and feelings of other individuals in the room. Thus, a demonstration of results, as opposed to techniques, was not, after all, going to be possible during our visit.

Backster then showed us videotapes of the split-screen results he had obtained in his "formal" experiments. The results consisted of 12 examples of apparent correlations between an emotional response and a deflection of the polygraph record. The 12 examples came from 7 sessions with 7 different subjects. Although the information is not given in his written report, it appears that each session lasted for approximately half an hour. During this time, the donor is engaged in conversation or watches videotapes of television programs. The sessions are not standardized or planned. Backster's intent, apparently, is to elicit spontaneous emotional responses from a subject during the session. He believes that a stimulus that evokes an emotional response in one subject will not necessarily do so in another subject.

In one example, the subject was a young man who was looking at an issue of *Playboy* magazine. The polygraph tracing began to display large deflections soon after he encountered a nude photograph of an attractive young woman. The large deflections continued for approximately two minutes; the tracing slowly settled down to normal activity after the magazine was closed. Soon after, the young man reached for the closed magazine, and the record reveals a single deflection at that point. In another example, the subject was a retired police lieutenant. When discussing his approaching retirement, he was asked a question about his wife's attitude toward having him "underfoot." A large deflection of the polygraph tracing occurred soon after this question was asked. When asked, the donor confirmed that he was emotionally aroused at that moment in the session (see Backster and White, 1985).

Cleve Backster and his supporters apparently believe that he has successfully demonstrated that detached oral leukocytes respond to the emotions of their donor even when separated by as much as several miles. They also believe that these results are reliable and replicable.

Critique of the Backster Experiment

What we have read and observed about Backster's procedures does not justify the claim he is making. His answers to our questions made it clear that he has not considered using the appropriate controls needed to ensure that the obtained "correlations" are real and due to the causes he has assumed. To make adequate physiological recordings from a

preparation of in vitro leukocytes and to demonstrate the correlation between emotional response and leukocyte activity requires experimental arrangements and procedures at a level of sophistication well beyond those we observed.

Committee members who are knowledgeable about the procedures and instrumentation of psychophysiological experiments expressed doubts about the adequacy of the setup to perform the tasks Backster has undertaken. Serious doubts were expressed about the possibility that the leukocytes were alive at the time of recording. Further doubts were expressed about the setup's ability to avoid contamination of the recording procedures by stray influences of various sorts. We do not discuss these drawbacks in detail here. We confine our discussion to Backster's method for establishing a correlation between the alleged activity of the detached leukocytes and the emotional state of the donor. When we consider how the existence of such correlations was established, we again see how inappropriate methodology can lead to very misleading conclusions.

Many problems exist with regard to Backster's procedures for detecting correlations. In trying to demonstrate a pattern of covariation between two records of behavior over time, one record is the tracing of amplified electrical activity coming from the electrodes and through the leads. Although this tracing can be quantified, Backster has apparently made no attempt to do so. Instead, he has relied on visual inspection of the polygraph record to pick out points at which the deflections of the pen from the baseline are noticeable. Although such subjective judgment is scientifically unacceptable, the deflections that he uses in his examples seem sufficiently marked that they probably can be considered to be real deviations from the baseline. At any rate, let us assume that responses on the polygraph record can be visually pinpointed with reasonable objectivity.

The deflections on the polygraph record are then compared with happenings on the concurrent videotaping of the conversation with the subject. Here we encounter very serious problems as to what constitutes an emotional response on this behavioral record. Backster believes he can identify categories of potentially emotionally arousing stimuli in the nonstandardized, qualitative, ongoing record of conversation. He then can determine if the subject was experiencing an emotional reaction to such a stimulus by simply replaying the record, pointing to the segment that corresponds to a place where the polygraph showed a deflection, and asking the subject if he or she recalls what was taking place at that moment as an emotionally arousing experience. If the subject agrees, this is said to confirm a "correlation" between the emotional state and the corresponding activity of the tracing.

Such a purely subjective determination of an emotional response opens

the process to a variety of known biases, many of them discussed in the paper prepared for the committee by Griffin (Appendix B). The literature on "illusory correlation" (Alloy and Tabachnik, 1984; Griffin paper) makes it clear how subjective expectations and cognitive biases can lead to false impressions of correlation. Backster's method of searching for correlations compounds these inevitable biases: he does not independently determine moments of emotional response in the subject's behavioral record and moments of polygraph deflections and then look for a match between the two. Instead, he apparently looks for polygraph deflections and then tries to determine if an emotional response can be found that occurred in the vicinity of the polygraph activity. In other words, the determination of the emotional response is done with full knowledge of the fact that a polygraph deflection has occurred.

Under such circumstances, we would expect processes of subjective validation to operate. In addition, the method of verifying the emotional response, by asking the subject to acknowledge that he or she was in fact experiencing such a state at the moment the polygraph record indicated a leukocyte response, is itself suspect. This is the sort of circumstance in which demand characteristics (i.e., responses determined by the presumed intent of the experimenters) are known to operate.

Good science dictates that the moments of emotional response should be determined independently of the moments of polygraph response. Both the experimenter and the subject must be blind to the polygraph record when determining the moments of emotional response. Only when the determination of events on the two records has been made independently of each other can the records be compared to determine if the emotional responses and the polygraph activity are correlated.

Illusory correlations occur because our subjective judgments of covariation tend to use only a portion of the relevant information and because we tend to bias observed events in terms of our expectations. In particular, intuitive judgments of covariation tend to focus only on the co-occurrence of treatment of interest and successful outcomes, ignoring times when the treatment co-occurred with unsuccessful outcomes. Backster uses only those examples from his records in which an emotional response co-occurs with a polygraph deflection; the 12 such examples from the 7 experimental series represent a very small fraction of the total data collected.

Not only is a sample of just 12 co-occurrences probably too small for estimating whether a true correlation exists, but it is also impossible from this information alone to estimate whether any correlation exists. All the data are needed for this purpose. Almost certainly, more than 12 polygraph deflections must have appeared in the total record. In the brief demonstration for the committee, both the control and the experimental series

yielded several deflections, so it is reasonable to assume that many more than 12 deflections were obtained in the complete record. It is likely that these unreported deflections were not preceded by any emotional responses.

Almost certainly, more than 12 emotional responses must have appeared in the total record. The point of conducting the sessions was to expose the subjects to a variety of emotional stimuli; therefore, it is essential to know the number of times that emotional responses occurred *without* the corresponding occurrence of polygraph responses. Finally, to determine correlation, it is essential to know the frequency of co-occurrence of the absence of emotional responses and the absence of polygraph responses. All this information is needed to determine whether the claimed correlation exists. All the data must be used. From these data, one can compare the proportion of times that an emotional response is followed by a polygraph response with the proportion of times that the absence of an emotional response is followed by a polygraph response. Only if these two proportions are significantly different from one another can we assume that the data provide evidence for a correlation between emotional response and leukocyte activity. The fact that Backster was able to find 12 examples of the co-occurrence between emotional response and polygraph deflection, even if these correspondences had come from double-blind matching, provides us with absolutely no information about whether a correlation exists.

The stronger claim would be, of course, not that a correlation exists, but that a causal connection exists between the subject's emotional states and the responses of the detached leukocytes. As Chapter 3 on evaluation indicates, such a causal explanation requires much more than the demonstration of correlation between two series. Because Backster did not use double-blind procedures to determine emotional responses, and because the procedures he did use are known to be just those that facilitate the occurrence of a variety of subjective biases, he may well have obtained a correlation between his two series. However, his procedures for finding such correlations are sufficiently flawed that we do not know if in fact the suspected (and presumably biased) correlation actually does exist in his data. The Backster experiment indicates that the best intentions combined with scientific instrumentation and polygraphic records cannot, in themselves, guarantee data of scientific quality.

Discussion of the Scientific Evidence

Both the parapsychologists cited in this report and the critics of parapsychology believe that the best contemporary experiments in parapsychology fall short of acceptable methodological standards. The critics

conclude that such data, based on methodologically flawed procedures, cannot justify any conclusions about psi. The parapsychologists argue that, while each experiment is individually flawed, when taken together they justify the conclusion that psi exists.

Palmer's conclusion in this regard is unique. Although he agrees that the data do not justify the conclusion that a paranormal phenomenon has been demonstrated, he argues that the data, with all their drawbacks, do justify the conclusion that an anomaly of some sort has been demonstrated. It is this purported demonstration of an anomaly that, according to Palmer, further justifies the claim that parapsychologists do have a subject matter. The awkward aspect of Palmer's position is that, without an adequate theory, there is no way to know that the anomaly "demonstrated" in one experiment is the same anomaly "demonstrated" in another; indeed, there is no limit to the possible causes of the anomaly in a given experiment. Without an adequate theory, there is no reason to assume that the various anomalies constitute a coherent or intelligibly related class of phenomena.

The committee distinguishes among three types of criticism that can be leveled at a given parapsychological finding. The first is what we might refer to as the smoking gun. This type of criticism asserts or strongly implies that the observed findings were due not to psi but to factor X. Such a claim puts the burden of proof on the critic. To back up such a claim, the critic must provide evidence that the results were in fact caused by X. Many of the bitterly contested feuds between critics and proponents have often been the result of the proponent's assuming, correctly or incorrectly, that this type of criticism was being made.

The second type of criticism can be referred to as the plausible alternative. In this case, the critic does not assert that the result *was* due to factor X, but instead asserts that the result *could have been* due to factor X. Such a stance also places a burden on the critic, but one not so stringent as the smoking gun assertion. The critic now has to make a plausible case for the possibility that factor X was sufficient to have caused the result. For example, optional stopping of an experiment on the part of a subject can bias the results, but the bias is a small one; it would be a mistake to assert that an outcome was due to optional stopping if the probability of the outcome is extremely low. Akers's critique, which was previously discussed, is an example based on the plausible alternative.

The third type of criticism is what we have called the dirty test tube. In this case, the critic does not claim that the results have been produced by some artifact, but instead points out that the results have been obtained under conditions that fail to meet generally accepted standards. The gist of this type of criticism is that test tubes should be clean when doing

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

careful and important scientific research. To the extent that the test tubes were dirty, it is suggested that the experiment was not carried out according to acceptable standards. Consequently, the results remain suspect even though the critic cannot demonstrate that the dirt in the test tubes was sufficient to have produced the outcome. Hyman's critique of the Ganzfeld psi research and Alcock's paper on remote viewing and random number generator research are examples of this type of criticism.

In the committee's view, it is in this latter sense, the dirty test tube sense, that the best parapsychological experiments fall short. We do not have a smoking gun, nor have we demonstrated a plausible alternative; but we imagine that even the parapsychological community must be concerned that their best experiments still fall far short of the methodological adequacy that they themselves profess.

Honorton and Hyman differ on whether to assign a flaw in randomization to a particular series of experiments. With Honorton's assignment, the studies with adequate randomization do not differ in significance of outcome from those with inadequate randomization. With Hyman's assignment, the experiments with inadequate randomization have significantly more successful outcomes than do those with adequate randomization. A simple disagreement on one experiment can thus make a huge difference as to whether we conclude that this flaw contributed or did not contribute to the observed outcomes. Several similar examples could be cited to illustrate the extreme sensitivity of this data base to slight changes in flaw assignments.

Even if Palmer is correct in asserting that in a particular case an anomaly has been demonstrated, serious problems remain. In astronomy and other sciences, an anomaly is a very precise and specifiable departure from a well-defined theoretical expectation. Neptune was discovered, for example, when Leverrier was able to specify not only that the orbit of Uranus departed from that expected by Newtonian theory, but also precisely in what way it departed from expectation. Nothing approaching such a specifiable anomaly has been claimed for parapsychology. A vague and unspecifiable departure from chance is a far cry from a well-described and systematic departure from a precise, theoretical equation. Leverrier's anomaly was consistent with only a very narrow range of possibilities. The sort of anomaly claimed for parapsychology is currently consistent with an almost infinite variety of possibilities, including artifacts of various kinds.

THE PROBLEM OF QUALITATIVE EVIDENCE

The committee continually encountered the distinction between qualitative and quantitative evidence for the existence of paranormal phenomena. Many proponents of the paranormal acknowledge such a difference in one way or another. Some realize that it is only quantitative evidence that will convince the scientific community. Although they themselves have relied on qualitative evidence for their own beliefs, they refer us to the RNG experiments of Robert Jahn or the remote viewing experiments at SRI as examples of supporting quantitative data.

Most proponents seem impatient with the request for scientific evidence. They have been convinced through their own experiences or the vivid testimonies of individuals whom they trust. Many argue that qualitative evidence can be as good as quantitative; indeed, they claim that in some circumstances it can be better.

The arguments for the superiority of qualitative evidence are based in many cases on such factors as ecological validity, conducive atmosphere, and holism. The ecological validity argument asserts that the artificial conditions required for laboratory experiments are so different from the natural settings in which paranormal phenomena typically occur that findings from such controlled studies are irrelevant. By removing the psychic from his or her natural domain or by arranging conditions to suit the needs of scientific observation, it is claimed, the scientist destroys the very phenomenon under question. The ecological validity argument is closely related to the other arguments. Proponents who emphasize the conducive atmosphere assert that the austere conditions of strict laboratory procedure create an atmosphere that is numbing or inimical to psychic functioning. Those who emphasize holism point out that the experimental procedures necessarily dissect and focus on restricted portions of a system. Such compartmentalization, it is claimed, makes it impossible to study the sorts of paranormal phenomena that operate only as a total system in a naturalistic context.

Qualitative Evidence and Subjective Biases

What is meant by qualitative evidence? Roughly, it means any sort of nonscientific evidence that proponents find personally convincing. Typically, it involves personally experiencing or witnessing the phenomenon. Less compelling, but still effective, is the testimony of friends or trusted acquaintances who have personally experienced it. Even individuals who are intellectually aware of the pitfalls of personal observation and testimony find it difficult, even impossible, to disregard the compelling quality of such evidence in the formation of their own beliefs.

A major parapsychologist admitted to one committee member that the scientific evidence did not justify concluding that psi exists. "As a trained scientist," he said, "I know quite well that by scientific criteria there is no evidence for the existence of psi. In fact, I have always argued with

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

my parapsychological colleagues that they are making a serious mistake in trying to get the scientific community to take their current evidence seriously. Before they do this, they first have to be able to collect the sort of repeatable and lawful data that constitute scientific evidence." This same parapsychologist then explained why, despite the current lack of evidence, he remained a parapsychologist. "When I was 16 I had some personal experiences of a psychic nature that were so compelling that I have no doubt that they were real. Yet, as a trained scientist, I know that my personal experiences and subjective convictions cannot and should not be the basis for asking others to believe me." This parapsychologist is unusual in that he makes the distinction within himself between beliefs that are subjectively compelling and beliefs that are scientifically justifiable. More typical is the proponent who, as a result of compelling personal experience, not only has no doubt about the reality of underlying paranormal cause, but also has no patience with the refusal of others to support that belief.

We see two problems regarding qualitative evidence. First, personal observation and testimony are subject to a variety of strong biases of which most of us are unaware. When such observations and testimony emerge from circumstances that are emotional and personal, the biases and distortions are greatly enhanced. Psychologists and others have found that the circumstances under which such evidence is obtained are just those that foster a variety of human biases and erroneous beliefs. Second, beliefs formed under such circumstances tend to carry a high degree of subjective certainty and often resist alteration by later, more reliable disconfirming data. Such beliefs become self-sealing, in that when new information comes along that would ordinarily contradict them, the believers find ways to turn the apparent contradictions into additional confirmation.

The committee asked Dale Griffin to describe many of the ways in which cognitive and social psychologists have documented that human subjective judgment can lead us astray. Griffin's paper emphasizes the cognitive biases termed *availability* and *representativeness*, but he also discusses motivational biases. Although most of these biases have been created under laboratory conditions, they are nonetheless quite powerful, and evidence has been mounting that, if anything, they are much more powerful in natural settings. Griffin points out that one vivid, concrete experience is usually sufficient to outweigh conclusions based on hundreds or thousands of cases based on abstract summary statistics. These and the other biases discussed by Griffin should make us wary of conclusions based on qualitative evidence.

Examples of Problematic Beliefs

In this section we discuss some examples of beliefs about paranormal phenomena that have been formed under conditions known to generate cognitive illusions and strong delusional beliefs. We attempt to make clear why we are skeptical of any evidence offered in support of the paranormal that does not strictly fulfill scientific criteria. We believe it is important to realize the power of such conditions to create strong but false beliefs.

In 1974 a group of distinguished physicists at the University of London observed renowned psychic Uri Geller apparently bend metallic objects and cause part of a crystal, encapsulated in a container, to disappear.

Impressed with what they saw, in 1975 these scientists contributed an article to *Nature* outlining their ideas about how to conduct successful parapsychological research (reprinted in Hasted et al., 1976). In their discussion they note that successful results depend on the relation among the participants and that phenomena are more likely to occur when all participants are in a relaxed state, all sincerely want the psychic to succeed, and "the experimental arrangement is aesthetically or imaginatively appealing to the person with apparent psychokinetic powers."

Hasted and his colleagues describe further desiderata. The psychic should be treated as one of the experimental team, contributing to an attitude of mutual trust and confidence that facilitates successful appearance of the allegedly paranormal effects. The slightest hint of suspicion on the part of the observers can stifle the occurrence of any phenomena. Observers should avoid looking for any particular outcome that interferes with the required relaxed state of mind and impedes paranormal powers. To help avoid the inhibiting effects of concentrated attention, participants should talk and think about matters irrelevant to the experiment at hand.

Acknowledging that these desiderata make it difficult to preclude trickery, Hasted and his colleagues express confidence that they can both create psi-conducive conditions and eliminate the possibility of being tricked (Hasted et al., 1976:194):

> It should be possible to design experimental arrangements which are beyond any reasonable possibility of trickery, and which magicians will generally acknowledge to be so. In the first stages of our work we did in fact present Mr. Geller with several such arrangements, but these proved aesthetically unappealing to him.

Although we may sympathize with the British physicists' desire to create conditions conducive to the appearance of genuine psychic powers, if such powers exist, we cannot fail to note the quandary that their efforts produce. In their quest for psi-conducive conditions, they have created guidelines that play into the hands of anyone intent on deceiving them.

The very conditions that are specified as being conducive to the appearance of paranormal phenomena are almost always precisely those that are conducive to the successful performance of conjuring tricks. One of the first rules the aspiring conjuror learns is never to announce in advance the specific outcome that he or she is going to produce. In this way onlookers will not know where and on what they should focus their attention and consequently will be less apt to detect the method by which the trick was accomplished. The authors' advice to avoid focusing on a predetermined outcome greatly facilitates the conjuror's task.

The insistence that the arrangements meet with the psychic's approval is by far the most devastating of these conditions. Geller will perform only if the conditions are "aesthetically pleasing." This amounts to giving the alleged psychic complete veto power over any situation in which he or she feels that success is not ensured. This in turn means that the psychic being tested, not the experimenters, is controlling the experiment. Surely the British physicists ought to realize the irony of their admission that all their experimental arrangements designed to preclude trickery turned out to be aesthetically unacceptable to Uri Geller.

Another example of beliefs generated in circumstances that are known to create cognitive illustions is macro-PK, which is practiced at spoon-bending, or PK, parties. The 15 or more participants in a PK party, who usually pay a fee to attend and bring their own silverware, are guided through various rituals and encouraged to believe that, by cooperating with the leader, they can achieve a mental state in which their spoons and forks will apparently soften and bend through the agency of their minds.

Since 1981, although thousands of participants have apparently bent metal objects successfully, not one scientifically documented case of paranormal metal bending has been presented to the scientific community. Yet participants in the PK parties are convinced that they have both witnessed and personally produced paranormal metal bending. Over and over again we have been told by participants that they know that metal became paranormally deformed in their presence. This situation gives the distinct impression that proponents of macro-PK, having consistently failed to produce scientific evidence, have forsaken the scientific method and undertaken a campaign to convince themselves and others on the basis of clearly nonscientific data based on personal experience and testimony obtained under emotionally charged conditions.

Consider the conditions that leaders and participants agree facilitate spoon bending. Efforts are made to exclude critics because, it is asserted, skepticism and attempts to make objective observations can hinder or prevent the phenomena from appearing. As Houck, the originator of the PK party, describes it, the objective is to create in the participants a

peak emotional experience (Houck, 1984). To this end, various exercises involving relaxation, guided imagery, concentration, and chanting are performed. The participants are encouraged to shout at the silverware and to "disconnect" by deliberately avoiding looking at what their hands are doing. They are encouraged to shout Bend! throughout the party. "To help with the release of that initial concentration, people are encouraged to jump up or scream that theirs is bending, so that others can observe." Houck makes it clear that the objective is to create a state of emotional chaos. "Shouting at the silverware has also been added as a means of helping to enhance the emotional level in a group. This procedure adds to the intensity of the command to bend and helps create pandemonium throughout the party."

A PK party obviously is not the ideal situation for obtaining reliable observations. The conditions are just those which psychologists and others have described as creating states of heightened suggestibility and implanting compelling beliefs that may be unrelated to reality. It is beliefs acquired in this fashion that seem to motivate persons who urge us to take macro-PK seriously. Complete absence of any scientific evidence does not discourage the proponents; they have acquired their beliefs under circumstances that instill zeal and subjective certainty. Unfortunately, it is just these circumstances that foster false beliefs.

Discussion of Qualitative Evidence

Our analysis of the evidence put before us indicates that even the most solidly based arguments for the existence of paranormal phenomena fall short of the currently accepted parapsychological standards. Even if the best evidence had been collected according to acceptable scientific standards, most proponents would have in fact remained convinced by personal experiences and data that clearly fall far short of scientific acceptability. We have looked at two examples to make clear why and in what ways such failures to meet acceptable standards render the corresponding arguments useless as evidence for the paranormal, even though they have created compelling and strongly held beliefs in those who have been exposed to them.

The examples illustrate how different ways of attempting to acquire evidence for paranormal phenomena can depart from adequate standards. These inadequacies become especially critical when we note that the conditions under which the alleged paranormal phenomena are supposed to occur are just those known to foster biases and false beliefs. The PK parties, while creating powerful beliefs in paranormal metal bending, clearly violate almost every principle for obtaining trustworthy data. These parties offer no standardization, no objective records, and no

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

controls against self-deception or the deliberate deception of others. All participants, including the leader, are encouraged to achieve a peak emotional state, and general chaos is encouraged.

The suggestions of a group of British physicists for testing alleged psychics are aimed at somehow combining the desire to keep the psychic from feeling inhibited with the desire to obtain evidence of acceptable scientific quality. The observers' zeal for making the psychic feel trusted produces conditions that make scientific observation impossible: observers are instructed to refrain from focusing attention on any expected result, and the experimental arrangement must be aesthetically acceptable to the psychic, a condition that in effect puts the psychic in control of the experiment.

The search for psi-conducive conditions is understandable. Parapsychological research, even at its best, has been continually frustrated by the lack of robust, lawful, and repeatable outcomes, yet parapsychologists have experienced phenomena or have encountered data that have convinced them of the reality of the paranormal. When they try to put such evidence before their critics, however, the phenomena have a habit of disappearing. If one fervently believes that the phenomena are real, then it becomes easy to imagine a variety of reasons why they are elusive and hard to produce on demand.

When proponents encounter a new phenomenon or psychic, they are strongly motivated to create conditions that will not drive the phenomenon away. The special atmosphere of PK parties and the suggestions of the British physicists are just two examples of attempts to generate psi-conducive conditions that also seem to be deception-conducive and bias-conducive.

CONCLUSIONS

In drawing conclusions from our review of evidence and other considerations related to psychic phenomena, we note that the large body of research completed to date does not present a clear picture. Overall, the experimental designs are of insufficient quality to arbitrate between the claims made for and against the existence of the phenomena. While the best research is of higher quality than many critics assume, the bulk of the work does not meet the standards necessary to contribute to the knowledge base of science. Definitive conclusions must depend on evidence derived from stronger research designs. The points below summarize key arguments in this chapter.

1. Although proponents of ESP have made sweeping claims, not only for its existence but also for its potential applications, an evaluation of the best available evidence does not justify such optimism. The strongest

claims have been made for remote viewing and the Ganzfeld experiments. The scientific case for remote viewing is based on a relatively small number of experiments, almost all of which have serious methodological defects. Although the first experiments of this type were begun in 1972, the existence of remote viewing still has not been established. Furthermore, although success rates varying from 30 to 60 percent have been claimed for the Ganzfeld experiments, the evidence remains problematic because all the experiments deviate in one or more respects from accepted scientific procedures. In the committee's view, the best scientific evidence does not justify the conclusion that ESP—that is, gathering information about objects or thoughts without the intervention of known sensory mechanisms—exists.

2. Nor does scientific evidence offer support for the existence of psychokinesis—that is, the influence of thoughts upon objects without the intervention of known physical processes. In the experiments using random number generators, the reported size of effects is very small, a hit rate of no more than 50.5 percent compared with the chance expectancy of 50 percent. Although analysis indicates that overall significance for the experiments, with their unusually large number of trials, is probably not due to a statistical fluke, virtually all the studies depart from good scientific practice in a variety of ways; furthermore, it is not clear that the pattern of results is consistent across laboratories. In the committee's view, any conclusions favoring the existence of an effect so small must at least await the results of experiments conducted according to more adequate protocols.

3. Should the Army be interested in evaluating further experiments, the following procedures are recommended: first, the Army and outside scientists should arrive at a common protocol; second, the research should be conducted according to that protocol by both proponents and skeptics; and third, attention should be given to the manipulability and practical application of any effects found. Even if psi phenomena are determined to exist in some sense, this does not guarantee that they will have any practical utility, let alone military applications. For this to be possible, the phenomena would have to obey causal laws and be manipulable.

4. The committee is aware of the discrepancy between the lack of scientific evidence and the strength of many individuals' beliefs in paranormal phenomena. This is a cause for concern. Historically, many of the the world's most prominent scientists have concluded that such phenomena exist and that they have been scientifically verified. Yet in just about all these cases, subsequent information has revealed that their convictions were misguided. We also are aware that many proponents believe that the scientific method may not be the only, or the most

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

appropriate, method for establishing the reality of paranormal phenomena. Unfortunately, the alternative methods that have been used to demonstrate the existence of the paranormal create just those conditions that psychologists have found enhance human tendencies toward self-deception and suggestibility. Concerns about making the experimental situation comfortable for the alleged psychic or conducive to paranormal phenomena frequently result in practices that also increase opportunities for deception and error.

SOURCES OF INFORMATION

Two of the military officers who briefed us during our first meeting urged the committee to give serious consideration to paranormal phenomena and related parapsychological techniques. They described a variety of such phenomena that they felt had military potential, either as threats to security or as aids to defense. Site visits to leading laboratories and a paper prepared for the committee also contributed to the bases for the committee's work. Briefings were given to committee members by Robert Jahn, Cleve Backster, Helmut Schmidt, members of the staff of the Stanford Research Institute, and the U.S. Army Laboratory Command in Adelphi, Maryland. The paper prepared by James Alcock provided detailed reviews of the available evidence on random event generators and remote viewing. In addition, the committee benefited from a thorough review conducted for the Army Research Institute by John Palmer and from its own review of recent articles in the *Journal of Parapsychology* and other relevant periodicals and handbooks.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

controls against self-deception or the deliberate deception of others. All participants, including the leader, are encouraged to achieve a peak emotional state, and general chaos is encouraged.

The suggestions of a group of British physicists for testing alleged psychics are aimed at somehow combining the desire to keep the psychic from feeling inhibited with the desire to obtain evidence of acceptable scientific quality. The observers' zeal for making the psychic feel trusted produces conditions that make scientific observation impossible: observers are instructed to refrain from focusing attention on any expected result, and the experimental arrangement must be aesthetically acceptable to the psychic, a condition that in effect puts the psychic in control of the experiment.

The search for psi-conducive conditions is understandable. Parapsychological research, even at its best, has been continually frustrated by the lack of robust, lawful, and repeatable outcomes, yet parapsychologists have experienced phenomena or have encountered data that have convinced them of the reality of the paranormal. When they try to put such evidence before their critics, however, the phenomena have a habit of disappearing. If one fervently believes that the phenomena are real, then it becomes easy to imagine a variety of reasons why they are elusive and hard to produce on demand.

When proponents encounter a new phenomenon or psychic, they are strongly motivated to create conditions that will not drive the phenomenon away. The special atmosphere of PK parties and the suggestions of the British physicists are just two examples of attempts to generate psi-conducive conditions that also seem to be deception-conducive and bias-conducive.

CONCLUSIONS

In drawing conclusions from our review of evidence and other considerations related to psychic phenomena, we note that the large body of research completed to date does not present a clear picture. Overall, the experimental designs are of insufficient quality to arbitrate between the claims made for and against the existence of the phenomena. While the best research is of higher quality than many critics assume, the bulk of the work does not meet the standards necessary to contribute to the knowledge base of science. Definitive conclusions must depend on evidence derived from stronger research designs. The points below summarize key arguments in this chapter.

1. Although proponents of ESP have made sweeping claims, not only for its existence but also for its potential applications, an evaluation of the best available evidence does not justify such optimism. The strongest

claims have been made for remote viewing and the Ganzfeld experiments. The scientific case for remote viewing is based on a relatively small number of experiments, almost all of which have serious methodological defects. Although the first experiments of this type were begun in 1972, the existence of remote viewing still has not been established. Furthermore, although success rates varying from 30 to 60 percent have been claimed for the Ganzfeld experiments, the evidence remains problematic because all the experiments deviate in one or more respects from accepted scientific procedures. In the committee's view, the best scientific evidence does not justify the conclusion that ESP—that is, gathering information about objects or thoughts without the intervention of known sensory mechanisms—exists.

2. Nor does scientific evidence offer support for the existence of psychokinesis—that is, the influence of thoughts upon objects without the intervention of known physical processes. In the experiments using random number generators, the reported size of effects is very small, a hit rate of no more than 50.5 percent compared with the chance expectancy of 50 percent. Although analysis indicates that overall significance for the experiments, with their unusually large number of trials, is probably not due to a statistical fluke, virtually all the studies depart from good scientific practice in a variety of ways; furthermore, it is not clear that the pattern of results is consistent across laboratories. In the committee's view, any conclusions favoring the existence of an effect so small must at least await the results of experiments conducted according to more adequate protocols.

3. Should the Army be interested in evaluating further experiments, the following procedures are recommended: first, the Army and outside scientists should arrive at a common protocol; second, the research should be conducted according to that protocol by both proponents and skeptics; and third, attention should be given to the manipulability and practical application of any effects found. Even if psi phenomena are determined to exist in some sense, this does not guarantee that they will have any practical utility, let alone military applications. For this to be possible, the phenomena would have to obey causal laws and be manipulable.

4. The committee is aware of the discrepancy between the lack of scientific evidence and the strength of many individuals' beliefs in paranormal phenomena. This is a cause for concern. Historically, many of the the world's most prominent scientists have concluded that such phenomena exist and that they have been scientifically verified. Yet in just about all these cases, subsequent information has revealed that their convictions were misguided. We also are aware that many proponents believe that the scientific method may not be the only, or the most

appropriate, method for establishing the reality of paranormal phenomena. Unfortunately, the alternative methods that have been used to demonstrate the existence of the paranormal create just those conditions that psychologists have found enhance human tendencies toward self-deception and suggestibility. Concerns about making the experimental situation comfortable for the alleged psychic or conducive to paranormal phenomena frequently result in practices that also increase opportunities for deception and error.

SOURCES OF INFORMATION

Two of the military officers who briefed us during our first meeting urged the committee to give serious consideration to paranormal phenomena and related parapsychological techniques. They described a variety of such phenomena that they felt had military potential, either as threats to security or as aids to defense. Site visits to leading laboratories and a paper prepared for the committee also contributed to the bases for the committee's work. Briefings were given to committee members by Robert Jahn, Cleve Backster, Helmut Schmidt, members of the staff of the Stanford Research Institute, and the U.S. Army Laboratory Command in Adelphi, Maryland. The paper prepared by James Alcock provided detailed reviews of the available evidence on random event generators and remote viewing. In addition, the committee benefited from a thorough review conducted for the Army Research Institute by John Palmer and from its own review of recent articles in the *Journal of Parapsychology* and other relevant periodicals and handbooks.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

NATIONAL ACADEMY PRESS • 2101 Constitution Avenue, NW • Washington, DC 20418

The National Academy of Sciences is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Frank Press is president of the National Academy of Sciences.

The National Academy of Engineering was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Robert M. White is president of the National Academy of Engineering.

The Institute of Medicine was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Samuel O. Thier is president of the Institute of Medicine.

The National Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Frank Press and Dr. Robert M. White are chairman and vice chairman, respectively, of the National Research Council.

Library of Congress Cataloging-in-Publication Data

Enhancing human performance : issues, theories, and techniques / Daniel Druckman and John A. Swets, editors.
p. cm.
"Committee on Techniques for the Enhancement of Human Performance, Commission on Behavioral and Social Sciences and Education, National Research Council."
Bibliography: p.
Includes index.
ISBN 0-309-03792-1. ISBN 0-309-03787-5 (soft)
1. Self-realization—Congresses. 2. Performance—Psychological aspects—Congresses. I. Druckman, Daniel, 1939– . II. Swets, John Arthur, 1928– . III. National Research Council (U.S.). Committee on Techniques for the Enhancement of Human Performance.
BF637.S4E56 1987
158—dc19 87-31233
CIP

COMMITTEE ON TECHNIQUES FOR THE ENHANCEMENT OF HUMAN PERFORMANCE

JOHN A. SWETS, *Chair*, Bolt Beranek and Newman Inc., Cambridge, Mass.
ROBERT A. BJORK, Department of Psychology, University of California, Los Angeles
THOMAS D. COOK, Department of Psychology, Northwestern University
GERALD C. DAVISON, Department of Psychology, University of Southern California
LLOYD G. HUMPHREYS, Department of Psychology, University of Illinois
RAY HYMAN, Department of Psychology, University of Oregon
DANIEL M. LANDERS, Department of Physical Education, Arizona State University
SANDRA A. MOBLEY, Director of Training and Development, The Wyatt Company, Washington, D.C.
LYMAN W. PORTER, Graduate School of Management, University of California, Irvine
MICHAEL I. POSNER, Department of Neurology, Washington University
WALTER SCHNEIDER, Department of Psychology, University of Pittsburgh
JEROME E. SINGER, Department of Medical Psychology, Uniformed Services University of Health Sciences, Bethesda, Md.
SALLY P. SPRINGER, Department of Psychology, State University of New York, Stony Brook
RICHARD F. THOMPSON, Department of Psychology, Stanford University

DANIEL DRUCKMAN, *Study Director*
JULIE A. KRAMAN, *Administrative Secretary*

Contents

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

Preface

The Army Research Institute in 1984 asked the National Academy of Sciences to form a committee to examine the potential value of certain techniques that had been proposed to enhance human performance. As a class, these techniques were viewed as extraordinary, in that they were developed outside the mainstream of the human sciences and were presented with strong claims for high effectiveness. The committee was also to recommend general policy and criteria for future evaluation of enhancement techniques by the Army.

The Committee on Techniques for the Enhancement of Human Performance first met in June 1985. The 14 members of the committee were appointed for their expertise in areas related to the techniques examined. The disciplines they represent include experimental, physiological, clinical, social, and industrial psychology and cognitive neuroscience; one member is a training program director from the private sector. During the next two years, the committee gathered six times, met in toto or in part on several occasions with various representatives of the Army, conducted interviews and site visits and sent subcommittees on several others, and commissioned 10 analytical and survey papers. The committee also examined a variety of materials, including state-of-the-art reviews of relevant literature, reports commissioned by the Army Research Institute, and unpublished documents provided by institutes, practitioners, and researchers. The report that follows describes the committee's activities, findings, and conclusions. Though cast largely in terms of the sponsor's setting, this report is relevant to other settings, for example, industry. The next few paragraphs present some background.

That the United States Army should be concerned to enhance the performance of its personnel is self-evident. We know that young volunteers must become not only soldiers who do well in battle but also technicians who skillfully operate and maintain complex equipment in peace and war. We are aware, moreover, that personal skills are not enough: individuals are heavily dependent on each other within small groups, and groups of various sizes must work very effectively together to permit survival and ensure success. And, of course, all must be ready to give peak performances in situations of great hardship, uncertainty, and stress. In the face of these staggering requirements, one must realize that turnover of personnel is high and that the training time available—to impart the necessary cognitive, physical, and social skills—is brief.

So it comes as no surprise that the Army is on the lookout for techniques that can help enhance human performance. The Army Research Institute is charged with seeking out and developing such techniques: it does so by employing researchers in the human sciences and by supporting appropriate research in universities and other public and private organizations. It focuses largely on promising new techniques as they appear in the mainstream of behavioral, physiological, and social research. However, given the pressures and given a view of mainstream research as slow, narrow, and insufficiently targeted, it also comes as no surprise that some influential officers and certain segments of the Army want to cast a broader net to snare promising enhancement techniques. To do this, they look beyond traditional research organizations and practices to what are viewed as extraordinary techniques. These techniques are thought possibly to provide such unusual benefits as accelerated learning, learning during sleep, superior performance through altered mental states, better management of behavior under stress, more effective ways of influencing other people, and so on. There is also an initiative within the Army to consider techniques based on paranormal phenomena, for example, extrasensory perception to view remote sites and psychokinesis to influence the operation of distant machines.

Along with these urgings to examine, to try, or to implement extraordinary techniques come difficult new problems for those in the Army responsible for evaluation, as well as for those in the Army responsible for personnel and training practices. One issue is that proponents of such techniques are usually not content with traditional evaluation procedures or scientific standards of evidence, often giving more weight to personal experience and testimony. Furthermore, a typical technique of this kind does not arise from the usual research traditions of experiments published in refereed journals and peer review of cumulated evidence, but rather appears full-blown as a package promoted by a commercial vendor. What does the Army Training and Doctrine Command or the base commander

do when the need is great, the package is ready, the claims are for miracles, some senior officers are vocally supportive, and the evaluation criteria are fluid? What do Army intelligence agencies do when the same conditions apply and other nations are said to be active in investigating paranormal effects?

The committee decided to assess a representative set of the techniques in question and resolved to address the surrounding issues in an open-minded and thorough way. We therefore divided ourselves into a number of subcommittees organized according to the behavioral processes addressed by the several techniques: accelerated learning, sleep learning, guided imagery, split-brain effects, stress management, biofeedback, influence strategies, group cohesion, and parapsychology. In addition, a subcommittee on evaluation issues was formed to examine practices and standards relevant to all the techniques. Each chapter of the report was prepared by the appropriate subcommittee, but interactions were frequent and so the report represents a collaborative effort of all the members.

Chapter 1 provides a context for the committee's task and the Army's interest in enhancing performance, characterizes some particular techniques, and introduces some general issues in evaluating them. Chapter 2 presents the committee's findings about the techniques examined and conclusions about appropriate evaluation procedures. Chapter 3 treats the relevant evaluation issues more systematically and presents the committee's philosophy of evaluation as it pertains to the matter at hand. Chapters 4 through 8 deal with particular techniques but are organized in terms of more general psychological processes. Chapter 9 considers parapsychological techniques.

The report concludes with six appendixes. Appendix A briefly summarizes the key elements of each enhancement technique. Appendix B lists the ten papers commissioned by the committee and their authors. Appendix C lists the members and activities of the subcommittees and also the activities of the committee as a whole. Appendix D lists key terms used in the research on particular techniques. Appendix E discusses the application of scientific research by the military. Appendix F contains biographical sketches of the committee members.

As committee chair, I am now in the pleasant position of recounting the several contributors to the total committee process, a process that went remarkably well. Definition and guidance for the committee's task came primarily from Edgar M. Johnson, director of the Army Research Institute. Administrative and technical liaison was ably provided by project monitor George Lawrence, who worked closely with the committee in its various activities. They were supported well by several senior Army officers, including Colonel William Darryl Henderson, Commander of the Army Research Institute; Major General John Crosby,

Assistant Deputy Chief of Staff for Personnel; and General Maxwell R. Thurman, Vice Chief of Staff. The committee met with members of a resource advisory group that included Lieutenant General Robert M. Elton, chair, Deputy Chief of Staff for Personnel; Lieutenant General Sidney T. Weinstein, Assistant Chief of Staff for Intelligence; Dr. Louis M. Cameron, Director of Army Research and Technology; Major General Maurice O. Edmunds, Commander of the Soldier Support Center; and Major General Philip K. Russell, Commander of the Medical Research and Development Command. Among the Army staff who were very helpful to the committee are Colonel John Alexander and Mr. Robert Kraus; the names of many others appear in Appendix C.

The committee's two consultants contributed special expertise: Paul Horwitz (of Bolt Beranek and Newman Inc.) joined the site visits of the subcommittee on parapsychology and advised on physical aspects of experiments in that area; James Schroeder (of Southwest Research Institute) attended the committee's meeting at Fort Benning, Georgia, and advised on the application of scientific research by the military (see Appendix E). The committee also received special expertise by commissioning papers. These papers and their authors are listed in Appendix B.

At the National Research Council, David Goslin, executive director of the Commission on Behavioral and Social Sciences and Education, once again provided wise counsel and support. Ira Hirsh, commission chair, and William Estes, also representing the commission, gave valuable advice and encouragement. Thomas Landauer, a member of the NRC's Committee on Human Factors, provided liaison in the areas of our committees' mutual interests. The reviewers of this report gave us a good measure of reinforcement along with helpful critiques. Eugenia Grohman, associate director for reports, lent experience and wisdom to this report. Special gratitude is extended to Christine McShane, the commission's editor: her skillful editing of the entire manuscript contributed substantially to its readability, and the coherence of the volume owes much to her suggestions for organizing the material. Julie Kraman, as administrative secretary to the committee, earned its considerable appreciation for setting up efficient meetings and for handling all manner of tasks graciously and smoothly.

Daniel Druckman, study director of the project, receives the committee's great appreciation for his intellectual contributions across the broad range of topics considered as well as for his logistic support. Working closely with the authors of chapters and commissioned papers, he provided an integration of the several contributions as well as much of the introductory and interstitial material. He also served on two subcommittees in areas of his expertise.

The ultimate debt of anyone who finds this report useful, and my large

personal debt, is to the members of the committee. As individuals, their capabilities are broad and deep. As a group, they gave generously and productively of their time, were always engaged, responded to every challenge, and, especially, showed an exceptional talent for reaching consensus in a collegial, advised, and efficient way.

JOHN A. SWETS, *Chair*
Committee on Techniques for the
Enhancement of Human Performance

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

PART I

Overview

PART I CONSISTS OF THREE CHAPTERS. Chapter 1 sets the stage for the report. It describes the committee's task, provides background on the Army's interest in enhancement techniques, characterizes specific techniques examined by the committee, and identifies the main issues in evaluating the relation between techniques and human performance. Chapter 2 presents the committee's findings and conclusions. We draw general conclusions about the process of consideration given to any technique and state specific findings and conclusions for each of the areas of human performance examined.

Chapter 3 presents the committee's philosophy of evaluation as it pertains to enhancement techniques. Some of the issues involved concern the conduct of basic research; others concern the conduct of field tests. With respect to basic research, issues include the plausibility of inferences about novel concepts, causation, alternative explanations of causal relations, and the generalizability of causal relations. With respect to field tests, a number of questions are of interest: Does the enhancement program meet genuine Army needs? Is the resulting program implementable, given program design and resources? Do unintended side effects limit utility? Is the program more cost-effective than its alternatives? These questions underscore the reality that evaluation research is largely a pragmatic activity influenced by the organizational context in which it occurs.

1

Introduction

THE COMMITTEE'S TASK

At the request of the U.S. Army Research Institute, the National Research Council formed a committee to assess the field of techniques that are claimed to enhance human performance. The Institute asked the Council to evaluate the claims made by proponents of selected existing techniques and to address two general additional questions: (1) What are the appropriate criteria for evaluating claims for such techniques in the future? (2) What research is needed to advance our understanding of performance enhancement in areas related to the proposed techniques? The objectives of the committee's study are to provide an authoritative assessment of these questions for policymakers in research and development who are consumers of the techniques, as well as to consider their possible applications to Army training.

Many of the techniques under consideration grew out of the human potential movement of the 1960s, including guided imagery, meditation, biofeedback, neurolinguistic programming, sleep learning, accelerated learning, split-brain learning, and various techniques to reduce stress and increase concentration. Many of these techniques have gained popularity over the past two decades, promoted by persons eager to provide answers to problems of human performance or to prosper from them. While often using the language of science to justify their approach, these promoters are for the most part not trained professionals in the social and behavioral sciences. Nonetheless, they do appeal to basic needs for human performance, and the Army, like many other institutions, is attracted to the prospect of cost-effective procedures that can improve performance.

These institutions must evaluate the effects of such procedures, however. Issues include the appropriateness of a quick-fix approach, the distinction between the impact of an experience and actual change, and the plausibility of evidence indicating that something is happening even if the effects are not reproducible or the benefits uncertain.

A more conservative atmosphere in the 1980s is reflected in the way techniques are advanced. Motivation in the 1980s may be primarily entrepreneurial, not ideological, as it was in the 1960s. Advocates focus on relating the techniques to specific tasks, such as marksmanship, foreign language acquisition, fine motor skills, sleep inducement, and even combat effectiveness. Some techniques are in fact rooted in a scientific literature. For these reasons the various techniques have attracted the interest of institutions that have rejected, and would probably continue to reject, countercultural trends in society. Indeed, much attention has been given to these techniques by industrial, government, and military policymakers, as well as by the general public. For this reason especially, it is important to address the issues surrounding the claims made for effectiveness.

Elaborate training programs have grown, nourished by their developers' enthusiasm and salesmanship in a social context receptive to quick cures. For many of these programs, success in the marketplace is used to justify their approaches. For others, more esoteric concepts, including the role of neurotransmitters, the physics of neuromuscular programming, brain wave patterns, hemispheric laterality, high-access memory storage, preferred sensory modalities, and low-gain innervation of muscles, are used to attempt to provide scientific justification for the claims. The chapters that follow evaluate the evidence and theories used to support the claims of several popular techniques. Before turning to these evaluations, however, we provide some background on the Army's interest in these techniques, as well as a discussion of issues surrounding enhanced performance and issues in evaluating the relation between techniques and performance.

THE ARMY'S NEEDS

The Army motto, "Be all that you can be," symbolizes the current ethos of the institution, an army of excellence. Emphasis is placed on attaining certain ideals, such as fearlessness, cunning, courage, one-shot effectiveness, fatigue reversal, and nighttime fighting capabilities. These ideals are assumed to be realizable through training, even if the most effective techniques have not as yet been identified. The culture of improvement is further reinforced by the dilemma created by an all-volunteer Army and the demands of complex new computer technologies. Many civilians enter military service with only the required minimum of

formal education; most of these volunteers enlist in the Army. For this reason, the Army's emphasis on skill training is well founded.

The importance of the human element in combat is recognized in the Army Science Board's 1983 report "Emerging Concepts in Human Technology," which phrases the issue in terms of high yield at relatively low investment. Human capital is considered to be the best potential source for growth in Army effectiveness, both in terms of return on investment and as a moral imperative "if we are to commit our soldiers to fight outnumbered and win." The technologies singled out in the report are those that can improve creativity and innovation, learning and training, motivation and cohesion, leadership and management, individual, crew, and unit fitness, soldier-machine interface, and the general productivity of the Army's human resources.

The Board's report largely bypasses issues of systematic evaluation of enhancement techniques within the Army context, while addressing mechanisms for integrating them with Army activities. Little concern is shown for adducing relevant criteria to determine whether implementation is feasible. The Army's ambitious goals, combined with a reluctance to deal with the complexities surrounding issues of human performance, make this institution potentially susceptible to a variety of claims made by technique developers. It would therefore seem prudent to devise criteria for evaluating those claims.

A SELLER'S MARKET

Techniques for enhancement of human performance have received much attention in the popular press. They have been actively promoted by entrepreneurs who sense a profitable market in self-improvement. The American Society for Training and Development "estimates that companies are spending an astounding $30 billion a year on formal courses and training programs for workers. And that's only the tip of the iceberg" (*Wall Street Journal,* August 5, 1986). They are also taken seriously by the U.S. military, who are at times accused of losing the "mind race" to the Soviets (see, for example, Anderson and Van Atta, *Washington Post,* July 17, 1985). The Army has shown particular interest in techniques that help people acquire, maintain, or improve such skills as classroom learning, communication and influence, creativity, and accuracy in the execution of tasks requiring motor skills. Those that are cost-effective and produce relatively rapid results are likely to receive the most attention, along with research breakthroughs that could be a basis for new training programs. What are these techniques? What claims are being made for them? Is there evidence that substantiates these claims?

Examples of techniques include biofeedback (information about internal

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

processes), Suggestive Accelerative Learning and Teaching Techniques (a package of methods geared primarily toward classroom learning), hemispheric synchronization (a machine-aided process based on assumptions about right brain–left brain activities), neurolinguistic programming (procedures for influencing another person), and Concentrix (a procedure used to improve concentration on specific targets). Also of interest to the Army are such processes as group cohesion and stress reduction, as well as the claims for sleep learning, peak performance, and parapsychology. Together, these techniques and processes cover the major types of skills—motor, cognitive, and social. Several of them are described here briefly, along with illustrative claims found in brochures and course material.

Suggestive Accelerative Learning and Teaching Techniques (SALTT) is an approach to training that employs a combination of physical relaxation, mental concentration, guided imagery, suggestive principles, and baroque music with the intent of improving classroom performance. Some applications have included language training, typing instruction, and high school science courses. Attempts have been made to evaluate these applications, and many of these evaluations are published in the *Journal of the Society for Accelerative Learning and Teaching* (Psychology Department, Iowa State University). The following is a sampling of claims made in brochures and convention announcements: "A proven method which has broad potential application in U.S. Army training"; "It will significantly reduce training time, improve memory of material learned and introduce behavioral changes that positively affect soldier performance—self-esteem, self-confidence, and mental discipline"; and "Most students will prove to themselves that they have learned a far greater amount of material per unit of time with a greater amount of pleasure than they have ever previously done."

Neurolinguistic programming (NLP) refers to a set of procedures developed to influence and change the behaviors and beliefs of a target person. Its goals are mostly therapeutic, but its proponents also advocate the use of the techniques in advertising, management, education, and interpersonal activities. A small research literature, published primarily in the *Journal of Counseling Psychology*, has developed. Practitioners can be trained and certified at various institutes, and the National Association for Neurolinguistic Programming distributes a newsletter to its membership, currently about 500 persons. Illustrative claims and testimonials found in advertising materials include: "[NLP] has evolved a unique technology which encompasses a set of specific techniques enabling you to produce well-defined results" and "NLP . . . is clear, easy to learn, and brilliant." A typical slogan is that found in a brochure from the Potomac Institutes, Silver Spring, Maryland: "The difference

that makes the difference, for education, management, psychotherapy, psychiatry, business, law, health care, and the arts."

Hemi-Sync™, which is short for hemispheric synchronization, is a technique that consists of presenting two tones slightly differing in frequency to separate ears with stereo headphones to produce binaural beats. The long-known result is a tone that waxes and wanes at a frequency equal to the difference between the original tones. Pioneered as an enhancement technique by Robert Monroe of the Monroe Institute of Applied Science in Faber, Virginia, the technique is based on the assumption of a frequency following response (FFR) in the human brain. The FFR refers to a correspondence between sound signals heard by the ear and electrical signals recorded by an electroencephalograph (EEG). It is claimed that, by altering sound patterns, it is possible to alter states of awareness. Stated applications are in the areas of language learning, stress management, reading skills, and creativity and problem solving. Claims of effectiveness stated in the Monroe Institute's brochure are wide-ranging, covering education (e.g., "77.8 percent of a class reported improvement in mental-motor skills"), health (early recuperation, lower blood pressure), psychotherapy (stress reduction, working with terminally ill patients, teaching autistic children), and sleep restorative training (e.g., "forty of forty-five insomniacs reported that one-month use of Hemi-Sync™ tapes was at least as effective as medication, without the drug side effects").

SyberVision® is a scripted videotape that presents an expert (e.g., a world-class athlete) repeatedly performing fundamental skills of his or her activity (e.g., golf) without verbal instructions. It is based loosely on principles of vicarious learning, guided imagery, and mental rehearsal. Developed and marketed by SyberVision Systems Inc., San Leandro, California, the package includes a cassette and instruction manual with an appendix on the "simple physics of neuro-muscular programming." The appendix presents a scientific rationale for the technique, for example, "the more you see and hear pure movement, the deeper it becomes imprinted in your nervous system . . . and the more likely you are to perform it as a conditioned reflex," and "The decomposition of what is seen and sensorily experienced into an electromagnetic wave form is accomplished by a complex mathematical operation (Fourier Transform) by the brain" (*Instruction Manual on Golf with Patty Sheehan*). Support for enhanced performance is, however, based on testimonials rather than experiments, for example, Killy on skiing, a Stanford tennis coach on tennis, Professional Golf Association members on golf, Peters (*In Search of Excellence*) on achievement, Salk on leadership, and a variety of corporate executives and educators on self-improvement. Claims range from sweeping statements (e.g., "We owe these two men a large debt of

gratitude") to rather precise statements (e.g., "In 47 days I have lost 25 pounds [191 to 166], yet I look like I lost 40") (in the United Airlines magazine, *Discoveries*). This technique involves a significant marketing effort that builds on users' willingness to be quoted and the use of acknowledged academic experts (e.g., Stanford neuropsychologist Karl Pribram), whose role in the program is advertised as being central.

Stress management techniques are procedures designed to alleviate anxiety or tension. Catering to an age of anxiety, self-help books, groups, and clinics on managing stress proliferate. A good example of the approach is the recent book by Charlesworth and Nathan (1982), which emphasizes fitness, nutrition, managing time, general life-styles and life-cycles, as well as strategies such as progressive relaxation, autogenic training, and image rehearsal. Appendixes provide the reader with home practice charts, a guide to self-help groups, and suggested books and recordings. The groups offer their members information, emotional support, and a sense of belonging. Often stress management procedures are combined with a number of other techniques into a single package. The promoters often emphasize the total package rather than particular techniques; the packages usually combine several processes that, when acting together, are thought to produce significant effects.

The Army's needs for techniques that can improve performance make it subject to the sorts of claims illustrated above. While they and other consumers can avoid the more obvious pitfalls, the proliferation of choices and products and the lack of scientific evidence allow marketplace criteria to become the bases for decisions. But there are exceptions. Some techniques have received the attention of the scientific community, and evidence is available to be used as criteria in such areas as biofeedback, guided imagery, sleep learning, cohesion, and even for some aspects of psychic phenomena and neurolinguistic programming.

The literature has alerted us, for example, to the distinction between the effects of biofeedback on fine motor skills and on stress, to the different effects of mental and physical rehearsal, to placebo and Hawthorne effects in stress research, to the priming and repetition effects of material presented during sleep, to some dysfunctions of group cohesion, to the difficulties of replicating experiments on extrasensory perception, and to the implausibility of specialized sensory modalities as postulated by NLP (see Appendix D for key terms). These findings make evident a complex relation between technique and performance.

IMPROVED PERFORMANCE: COMPLEX ISSUES, SIMPLE SOLUTIONS

The research literature in such traditional areas of experimental psychology as learning, perception, sensation, and motivation suggests

complex relations between interventions and improved performance. Many technique promoters appear to pay little attention to this literature, preferring an alternative route to invention: rather than derive a procedure from appropriate scientific literature, they create techniques from personal experiences, sudden insights, or informal observation of "what works." Science may enter the process after the technique is developed and used, for example, to legitimize its use or to endorse methods for evaluation. Research follows rather than precedes the invention. This sequence increases the likelihood that important considerations will be missed. We highlight some of these considerations in this section.

The lack of easy avenues to improved performance may well be due to the complexity of the behavior in question. One definition of skills emphasizes the importance of the coordination of behavior: "A skilled response . . . means one in which receptor-effector-feedback processes are highly organized, both spatially and temporally. The central problem for the study of skill learning is how such organizations or patterning comes about" (Fitts, 1964:244). This definition implies that skill learning involves an orchestration of diverse processes, making the topic an interesting one to various subfields of psychology. It also makes evident a number of unresolved issues, including whether different skills are learned and retained in different ways. The research findings obtained in this literature contribute to our understanding of the necessary, if not sufficient, conditions for improved performance.

Research on skill acquisition addresses such basic questions as What are the stages of learning? and What is learned? Distinctions made between short-term and long-term memory storage and between schemas and details have contributed to our understanding of basic processes (see Welford, 1976). Other questions have more direct consequences for application: for example, what contributes to the acquisition and maintenance of skills? How can the adverse effects of stress, fatigue, and monotony be avoided? These questions are the basis for programs of research that can be divided into several parts, each defined in terms of empirical issues (Irion, 1969; see also the other chapters in Bilodeau and Bilodeau, 1969). Some examples of empirical issues are practice effects (differences due to distributed versus massed practice, long versus short rest periods, short versus long sessions), the whole-part problem (differences due to learning a task as a whole versus learning it by its constituent elements), feedback (differences due to delays in receiving knowledge of results and to type of information during the delay period), retention (differences due to whether the the task is motor or verbal), and transfer of training.

These and related considerations suggest that skill learning is an incremental process likely to differ from one type of skill to another. Whether intending to enhance motor, verbal, problem-solving, or social

performances, technique designers can ill afford to ignore these lessons from the experimental literature on skill acquisition and maintenance. It is also the case, however, that the agenda of unexplored issues is much larger than the accomplishments to date, and this is recognized particularly in the rapidly growing field of cognitive psychology, in which the "information-processing revolution" is just beginning.

Practical applications are, however, not automatic. Many excellent applications do not spring from basic science; some are the result of craft and experience. More important perhaps are the indirect contributions made in both directions—from basic to applied and vice versa. A systematic approach taken in both domains serves to vitalize each, as when applied investigations reveal new phenomena that need explanation or when a new package incorporates basic principles discovered originally in the laboratory. Such an approach is likely to facilitate the design of appropriate techniques for skill acquisition. At issue is whether a particular technique can produce and sustain desired changes.

One conclusion from the research accumulated to date is that effective interventions are those that are continuous and self-regulating and take account of both context and person (see, for example, Lerner, 1984). Particularly relevant is the difference between short-term and long-term changes. Effects obtained by many techniques for performance enhancement may be short-term in their effects. This distinction is made by Back (1973, 1987) in his evaluation of the sensitivity training movement. The changes observed by sensitivity trainers and documented by evaluators may well reflect the impact of the experience per se. Such situation effects are unlikely to be sustained in different environments, an observation supported by the literatures in both developmental and social psychology (Druckman, 1971; Frederiksen, 1972). These literatures caution against hasty generalizations from observed, situation-specific effects; they also explain why long-term effects may be difficult to produce with brief exposures to "treatments." Like the sensitivity trainers of the 1960s and 1970s, many of the promoters (and consumers) of the 1980s pay little attention to issues of causality and intrinsic motivation, preferring instead to dwell on single dimensions of treatments or to offer a mixed package constructed in arbitrary ways and producing diffuse effects that reflect the experience.

The issue of expected benefits from techniques provides a bridge between research and application. Research can be designed to evaluate techniques, as well as to discover possible unintended side effects. Indeed, a research literature has developed in some of the areas examined in this book, namely biofeedback, stress, and guided imagery. For many other techniques, however, a relevant body of research does not exist; this lack applies to some of the techniques examined by the committee,

as well as to those yet to appear on the market. It is these techniques that present a problem for us as evaluators. Evaluation without data is difficult, but not impossible. Our approach is to place the techniques into broader categories corresponding to the key processes being influenced, for example, learning, motor skills, and influence. By so doing, the claims can be evaluated within the frameworks of existing theories and methodologies. They can also be judged against results obtained in related areas. This approach serves as the organizing theme for the chapters that follow.

EVALUATING THE TECHNIQUES

Evaluations properly hinge on answers to a standard set of questions proposed in a paper entitled "Evaluating Human Technologies: What Questions Should We Ask?" by Hegge, Tyner, and Genser (1983) at the Walter Reed Army Institute for Research:

- What changes will the technique produce?
- What evidence supports the claims for the technique?
- What theories stand behind the technique?
- Who will be able to use the technique?
- What are the implications of the technique for Army operations?
- How does the technique fit with Army philosophy?
- What are the cost-benefit factors?

These questions served as guidelines for the committee's evaluations. Appendix A is a summary description of each technique, organized along the lines of the Hegge, Tyner, and Genser questions, covering theory, research, and application. For many of the categories, however, the desired information is either too limited to be useful or simply not available; in such cases we have considered other strategies for evaluation.

The committee faced a number of difficulties in evaluation that stem from recurrent problems posed by the technologies. One is the tendency for some promoters (and consumers) to rely primarily on testimonials or anecdotal evidence as a basis for application. Another is a general lack of strong research designs to provide evidence of effects. These problems are considered also in the context of specific techniques discussed in the chapters of Parts II and III.

Practitioners of techniques often emphasize the value of personal or clinical experience and marketplace popularity as bases for judging the techniques. They are generally less inclined to seek research evidence or to support research evaluation programs. These attitudes may be related to the fact that few practitioners are trained as researchers. For some it is sufficient to let others do the research. For others, research is

viewed, in varying degrees, as a threat to their product. At one extreme, research is regarded as a debunking enterprise, engaged in by scientists who have little interest in providing human services. At another extreme, the problem is one of educating the researchers in nuance, context, and a clinical approach that emphasizes adapting techniques to changed situations and client tastes. The result is a gap in communication epitomized by two cultures—scientists searching for evidence and practitioners seeking effects and cures. A step toward bridging the gap would consist of mutual education through joint ventures. These ventures would expose scientists to the goals (and motives) of practitioners and would also make practitioners aware of the general analytical approaches used by scientists.

Experimentation is an appropriate vehicle for evaluating performance-enhancing techniques; the problem is usually defined in terms of effects of techniques (procedures) on performance (behaviors). It is also appropriate at an earlier stage in the process, when products are being developed. Products evolve in a kind of trial-and-error fashion similar in many respects to scientific discoveries. One model for integrating research with product development is engineering research and development (R&D). A strenuous applied research effort accompanies the development process in many firms, as does a quality-control program designed to evaluate products both during development and after they have been placed on the market. With a few exceptions, this model has not been adopted by firms or institutions in the field of performance enhancement.

Experimental evidence has accumulated in some areas related to techniques. Although not linked specifically to product development in the manner of an R&D operation, this work does address the question, What evidence supports the claims for the technique? In fact, so strong is the experimental tradition in some areas that a body of work has developed programmatically within a generally accepted paradigm (e.g., guided imagery). The benefits of a long research tradition can be seen in these areas. Meta-analyses have been performed and can be used as a basis for evaluation. For other areas, we are presented with the prospect of relying on scattered experiments or using other criteria as a basis for evaluation, or both (see Appendix A for summaries of the state of the science in each of the areas).

However, the benefits of experimental evidence derive primarily from the general approach rather than from the particular experiments. This idea is captured by Kelman, who noted that "an experimental finding . . . cannot very meaningfully stand by itself. Its contribution to knowledge hinges on the conceptual thinking that has produced it and into which it is subsequently fed back" (1968:161). We emphasize here the contribution

of an analytical approach to thinking about behavior, as distinct from the establishment of laws about psychological processes. It is the cumulation of a series of experiments that winnows out the useful parts of treatments or techniques. It is the self-correcting progression of new experiments that refines treatments, saving those that work and discarding those that do not (or that work only under very restricted conditions). This process contributes equally well to the goals of theory development and product development.

Other evaluation criteria elucidated by Hegge, Tyner, and Genser (1983) include theories, uses, and implications for Army operations and philosophy. A problem with these criteria is that they tend to be vague and somewhat idiosyncratic, making it difficult to propose general categories on which most people would agree. Without precisely defined categories for judging techniques, it is difficult to address issues of transfer of performance from one situation to another or to evaluate newly emerging techniques. A similar problem exists with respect to developing taxonomies in broadly defined fields: there is little agreement on a set of categories for the fields of human learning, performance, motivation, perception, and social and organizational processes. More mature subdisciplines provide an empirical basis for taxonomies, allowing for more tightly constructed systems of tasks and situations: for example, rote learning, short-term memory, concept learning, problem solving, work motivation, and team functions (see Fleishman and Quaintance, 1984). An advantage of such systems is that they capture rather precise relationships between task and performance.

This discussion serves only to introduce the issues and identifies several themes that receive more detailed attention in the chapters to follow. First, any evaluation must take into account the status of the available evidence. Confidence placed in judgments about a technique should be based on the quality of the evidence produced by researchers. Second, the evaluator cannot afford to rely exclusively on a single criterion for judging effectiveness. Theoretical and applied issues are also important, as are considerations of values served or violated by use of the technique. Third, technique development issues are not isolated from research or analytical issues. Each step in the process of product design can be regarded as an empirical issue; decisions made about procedures and packaging can be the result of experimental outcomes. Fourth, the subject of enhancing human performance is not new. It has been a topic of interest for centuries and an area of scientific work for several decades. The literatures on learning and skill acquisition should be consulted by developers, and insights derived from these literatures should be used in product design.

These themes are woven throughout the discussions of specific techniques. Each chapter discusses relevant literature, describes the specific techniques, points to directions for further research when appropriate, and notes possible applications in military and industrial settings. Despite the common coverage, however, each chapter is also unique in that each is tailored to the particular problems associated with its focus.

2

Findings and Conclusions

The committee's first major task was to evaluate the existing scientific evidence for a wide range of techniques that have been proposed to enhance human performance. This evaluation was intended by our Army sponsors to suggest guidelines for decision making on Army research and training programs. In our evaluation we draw conclusions with respect to whether more basic or applied research is warranted, whether training programs could benefit from new findings or procedures, and what, in particular, might be worth monitoring for potential breakthroughs of use to the Army. In many of the areas examined it appears feasible to pursue carefully designed programs that build on basic research; however, such programs should be monitored closely.

The committee's second major task was to develop general guidelines for evaluating newly proposed techniques and their potential application. We are aware that the use of basic and applied research in decision making is a complex issue. Although payoffs from basic research can often be realized in the long run, the value of research findings to the Army depends on developing a way of putting them into practice. With regard to applied or evaluation research, further complexities are evident: multiple, sometimes conflicting, criteria must be satisfied at each of several stages in the evaluation process, from assessing a pilot program to implementing the program in an appropriate setting. Another problem is that of choosing among alternative techniques when none of them has been subjected to a systematic evaluation. In the absence of evaluation studies, the Army needs guidelines for selecting packages and vendors.

The committee's evaluation has produced several answers to questions

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

of how best to improve performance in specific areas. On the positive side, we learned about the possibilities of priming future learning by presenting material during certain stages of sleep, of improving learning by integrating certain instructional elements, of improving skilled performance through certain combinations of mental and physical practice, of reducing stress by providing information that increases the sense of control, of exerting influence by employing certain communication strategies, and of maximizing group performance by taking advantage of organizational cultures to transmit values. On the negative side, we discovered a lack of supporting evidence for such techniques as visual training exercises as enhancers of performance, hemispheric synchronization, and neurolinguistic programming; a lack of scientific justification for the parapsychological phenomena considered; some potentially negative effects of group cohesion; and ambiguous evidence for the effectiveness of the suggestive accelerative learning package.

The remainder of this chapter presents the committee's findings and conclusions, which are presented in two parts: general conclusions regarding the process of evaluating any technique being considered by the Army and specific findings and conclusions for each of the areas of human performance examined. Whenever appropriate, we make recommendations for research, evaluation, and practice.

GENERAL CONCLUSIONS

The committee suggests that the Army move vigorously, yet carefully and systematically, to implement techniques that can be shown to enhance performance in military settings. Such an effort would be timely because of recent developments in the relevant research areas. Moreover, the payoff is likely to be very high if techniques are selected judiciously. Although the desire for dramatic improvements in performance makes some extraordinary techniques attractive, techniques drawn from mainstream research in relevant areas of performance may be more effective. The Army's concern for enhancing human performance and its substantial resources for evaluating techniques place it in a favorable position to take advantage of developments. The Army might also consider the possibilities of transferring its findings to the civilian sector.

Collectively, the committee's conclusions call for the adoption of scientifically sound evaluation procedures; however, these procedures must be adapted to institutional needs and must take into account problems of implementation. We summarize these considerations below.

Scientific Evidence

Techniques and commercial packages proposed for consideration by the Army should be shown to be effective by adequate scientific evidence

or compelling theoretical argument, or both. A technique's utility should be judged in relation to alternatives designed for similar purposes, and the estimated utility should be of significant magnitude. Specific stages of analysis can be incorporated in pilot or field testing, and such testing should be carried out by investigators who are independent of the technique's originators or promoters.

Testimonials as Evidence

Personal experiences and testimonials cited on behalf of a technique are not regarded as an acceptable alternative to rigorous scientific evidence. Even when they have high face validity, such personal beliefs are not trustworthy as evidence. They often fail to consider the full range of factors that may be responsible for an observed effect. Personal versions of reality, which are essentially private, are especially antithetical to science, which is a fundamentally public enterprise. Of course, a caution about testimonials should not be confused with a lack of openness to new and unusual ideas. Such openness is consistent with the requirement that the evidential criteria of science be satisfied.

The subject of testimonials as evidence has received considerable attention in recent research on how people arrive at their beliefs. These studies indicate that many sources of bias operate and that they can lead to personal knowledge that is invalid despite its often being associated with high levels of conviction. The committee recommends that this research be disseminated, as appropriate, in the Army. It may then be applied whenever testimony is used as the primary evidence to promote an enhancement technique.

Conditions for Implementation

Two kinds of evidence should be sought to support decisions to implement a technique: successful field tests and an analysis of implementability. It would also be useful to analyze the impact of the technique or package on the larger system in which it is to be embedded. These analyses would aid in explaining why the procedures are necessary and why certain consequences are expected. In general, any description of what a technique accomplishes should be accompanied by an explanation of *why* it accomplishes what it does. Such an explanation would provide a more fundamental understanding of processes affected by exposure to the technique and permit optimal implementation.

Rational Decision Making

The considerations that must be entertained in selecting a technique for practical use in a military setting are different from the considerations

needed to verify the existence of an enhancement effect in a scientific setting. For example, the benefits of correct decisions and the costs of incorrect decisions, that is, the risk calculus, may differ in the two settings. Furthermore, what is viewed as a timely decision will also differ. The specific differences as they apply to particular decisions should be made explicit.

Mechanisms for Advice

It would be useful to provide valid information about useful techniques to Army commanders and other interested staff on a regular basis. Special consideration should be given to ways in which technique-related information can be transferred from scientists to practitioners. The characteristics of a transfer agent could be defined, and such a position might be established within an appropriate office.

The committee recommends that the Army Research Institute formalize the ways in which it receives and provides advice about specific techniques. A committee to review experimental designs and statistical analyses could be convened to improve the evaluation of techniques. Special and standing committees could also be used to make program recommendations and to review proposals for intramural and extramural research.

Bidding Procedures

Purchase by the Army of a commercial enhancement package should take place within the context of a set of well-defined procedures. The committee recommends that an open-bid procedure be followed, based on a full presentation of the Army's stated objectives. This would encourage competitive evaluation of techniques. The following information, presented in a standard format, should be required: the objectives of the technique, a description of its procedures, evidence that it produces the claimed effects, and the vendor's record of past achievements in relevant areas.

Lack of professional training and research experience in human performance by a designer or advocate should not preclude consideration of the proposed package; it should, however, signal the need for a more stringent analysis by the Army.

SPECIFIC FINDINGS AND CONCLUSIONS

We present below findings and conclusions for each of the areas investigated. Some statements take the form of suggested actions based on what we know; others consist of suggestions for more work or for research that has not yet been done.

Learning During Sleep

1. The committee finds no evidence to suggest that learning occurs during verified sleep (confirmed as such by electrical recordings of brain activity). However, waking perception and interpretation of verbal material could well be altered by presenting that material during the lighter stages of sleep. We conclude that the existence and degree of learning and recall of materials presented during sleep should be examined again as a basic research problem.

2. Pending further research results, the committee concludes that possible Army applications of learning during sleep deserve a second look. Findings that suggest the possibility of state-dependent learning and retention (i.e., better recall of material when learned in the same physiological and mental state) may be applicable to fatigued soldiers. Furthermore, even presentations of material that disrupt normal sleep may be cost-effective, as may presentations that coincide with stages of light sleep.

Accelerated Learning

1. Many studies have found that effective instruction is the result of such factors as the quality of instruction, practice or study time, motivation of the learner, and the matching of the training regimen to the job demands. Programs that integrate all these factors would be desirable. We recommend that the Army examine the costs, effectiveness, and longevity of training benefits to be derived from such programs and compare them with established Army procedures.

2. The committee finds little scientific evidence that so-called superlearning programs, such as Suggestive Accelerative Learning and Teaching Techniques, derive their instructional benefits from elements outside the mainstream of research and practice. We observe, however, that these programs do integrate well-known instructional, motivational, and practice elements in a manner that is generally not present in most scientific studies.

3. We find that scientifically supported procedures for enhancing skills are not being sufficiently used in training programs and make two recommendations to remedy this problem. First, the basic research literature should be monitored to identify procedures verified by laboratory tests to increase instructional effectiveness. Second, additional basic

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

research should be supported to expand the understanding of skill acquisition for both noncombat and combat activities.

4. We conclude that the Army training system provides a unique opportunity for cohort testing of training regimens. The Army is in a position to create laboratory classroom environments in which competing training procedures can be scientifically evaluated.

5. The committee recommends that the Army investigate expert teacher programs by identifying and evaluating particularly effective programs within the Army. In addition, transferable elements of effective instruction can be reported to the larger instructional community.

Improving Motor Skills

1. The committee concludes that mental practice is effective in enhancing the performance of motor skills. This conclusion suggests further work in two directions: (1) evaluation studies of motor skills used in the Army and (2) research designed to determine the combination of mental and physical practice that, on average, would best enhance skill acquisition and maintenance, taking into account both time and cost.

2. The committee concludes that programs purporting to enhance cognitive and behavioral skills by improving visual concentration have not been shown to be effective to date. In our judgment, these programs are not worth further evaluation at this time.

3. The committee concludes that existing data do not establish the generality of observed effects from programs that train visual capabilities to increase performance.

4. Similarly, the committee concludes that the effects of biofeedback on skilled performance remain to be determined.

5. The committee recommends additional research to establish the potential of these techniques in the domain of specific skilled performances.

Altering Mental States

1. Time did not allow the committee to explore the evidence for a wide variety of specific methods for relating mental states to changes in performance. Such methods include forms of self-induced hypnotic states and peak performance resulting from high levels of focused concentration and meditation. We recommend that reviews of the literature in these areas be undertaken to ascertain whether any practical results might be obtained by the use of such methods.

2. The committee finds that, while the study of mental computations in language and imagery has progressed in recent years, the effort to understand how such computations are modulated by energetic factors

such as arousal, stress, emotion, and high levels of sustained concentration has not been fully developed. For example, the claims that certain mental states produce general improvements in performance derive from the idea, supported by research, that arousal affects mental computations and that there ought to be an optimal level of arousal for the performance of such computations. We recommend this as an important area for investment of basic research funds.

3. The committee's review of the appropriate literature refutes claims that link differential use of the brain hemispheres to performance. Further evaluation of these claims depends on developing valid and reliable measures of hemispheric involvement.

4. The committee finds no scientifically acceptable evidence to support the claimed effects of techniques intended to integrate hemispheric activity, for example, Hemi-Sync®. Attempts to increase information-processing capacity by presenting material separately to the two hemispheres do not appear to be useful. We conclude that such techniques should be considered further by the Army only if scientific evidence is provided to and evaluated by the Army Research Institute.

Stress Management

1. Existing data indicate that stress is reduced by giving an individual as much knowledge and understanding as possible regarding future events. In addition, giving the individual a sense of control is effective. On the basis of these findings, the committee recommends a systematic program of research and development that would address three questions: (1) How relevant is this finding for stress reduction in the Army? (2) To what extent does stress reduction realized in training transfer to combat situations? (3) What are the limitations on providing knowledge and understanding of future events and a sense of control in the Army setting? Pending the outcome of this research, we suggest that consideration be given to including the material in training programs for company grade, field grade, command, and staff officers.

2. We find that, while biofeedback can achieve a reduction of muscle tension, it does not reduce stress effectively. It is therefore not a promising research topic in that respect. We recommend that funding be directed toward investigation of more promising stress management procedures.

3. We recommend that information be gathered on the costs of stress in terms of organ breakdown, loss of efficiency, and loss of time. This information would have implications for training programs.

Influence Strategies

1. The committee finds no scientific evidence to support the claim that neurolinguistic programming is an effective strategy for exerting influence.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

We advise that further Army study of this aspect of NLP be made only in comparison with other techniques.

2. There are no existing evaluations of NLP as a model of expert performance. We conclude that further investigation of such models may be worthwhile and suggest that NLP be examined in comparison with several other techniques.

3. Concerning the process of technology transfer, we recommend that studies be conducted to develop training regimens for those who train others to wield social influence. The large literature on this topic in social psychology would provide a basis for such packages.

Group Cohesion

1. We find few scientific studies that address the possible relationship between group cohesion and performance; however, such a relationship may well be found with more extensive research. There is a need for research to consider the possibility of negative effects from inducing cohesion and methods of avoiding such effects. The committee recommends continued study of cohesion and related group processes.

2. We are favorably impressed with the evaluation studies of the Army's COHORT system. We endorse the investigators' plan to proceed beyond measures of attitudes to measures of group performance.

3. We recommend that the Army, as well as independent investigators, study the possible impacts of cohesion beyond the COHORT system, for example, on intergroup performance.

Parapsychology

1. The committee finds no scientific justification from research conducted over a period of 130 years for the existence of parapsychological phenomena. It therefore concludes that there is no reason for direct involvement by the Army at this time. We do recommend, however, that research in certain areas be monitored, including work by the Soviets and the best work in the United States. The latter includes that being done at Princeton University by Robert Jahn; at Maimonides Medical Center in Brooklyn by Charles Honorton, now in Princeton; at San Antonio by Helmut Schmidt; and at the Stanford Research Institute by Edward May. Monitoring could be enhanced by site visits and by expert advice from both proponents and skeptics. The research areas included would be psychokinesis with random event generators and Ganzfeld effects.

2. One possible result of the monitoring mentioned above is the proposal

of specific studies. In that situation the committee recommends the following procedures: first, the Army and outside scientists should arrive at a common protocol; second, the research should be conducted according to that protocol by both proponents and skeptics; and third, attention should be given in such research to the manipulability and practical application of any effects found to exist.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

3

Evaluation Issues

Implementation of an enhancement technique, in the committee's view, should depend on two general kinds, or levels, of evaluation. The first examines primarily the scientific justification for the effectiveness of the technique and the potential of the technique for improving performance in practice. The second kind examines field tests of a pilot program incorporating the technique to determine how feasible it is and to what extent it brings about effects that Army officials consider useful.

Convincing scientific justification can come only from basic research, that is, from carefully controlled studies that usually take place in laboratory settings and that preferably are related to a body of theory. Such research can provide evidence for the existence of the causal effect on which a technique is based and can help explain, or indicate a mechanism for, the effect. Analysis in connection with basic research should go beyond scientific justification to operational potential and likely cost-effectiveness. Only field tests can assess a program's actual operations and effects, however, and for such tests a broader array of evaluative criteria are needed, related primarily to the technique's utility.

Because strong claims of support from basic research have been made for some of the techniques the committee examined, we review here what it takes to justify a scientific claim, specifically, we review some standards for evaluating basic research. We then examine in more detail some standards for evaluating field tests of pilot programs. In the third section of this chapter, we set forth briefly some of our impressions of how the Army now manages the solicitation and evaluation of new performance-enhancing techniques. This chapter concludes with a note

on informal, qualitative approaches to evaluation, which are sometimes suggested as alternatives to basic research and field tests.

This chapter does not aspire to a comprehensive treatment of evaluation issues, and it barely touches on research methods. Articles, journals, books, and handbooks testify to the scope and complexity of this burgeoning field (e.g., Barber, 1976; Cook and Campbell, 1979). Our objective here is to highlight the topics that have impressed us as most germane. The various sources just mentioned would need to be consulted for even a minimal elaboration of these topics, and other committees would be required if recipes for evaluation of the Army's enhancement programs were sought as extensions of our work. Still, we believe this chapter will help the Army set general evaluation standards.

STANDARDS FOR EVALUATING BASIC RESEARCH

The purpose of basic research is to permit inferences to be drawn in accordance with scientific standards, including inferences about novel concepts, about causation, about alternative explanations of causal relations, and about the generalizability of causal relations.

For novel concepts, evidence must be gathered that both the purported enhancement technique and the relevant performance have been (1) defined in a way to highlight their critical elements, (2) differentiated from related variables that might bring about similar effects, and (3) put into operation (manipulated or measured) in ways that include the critical parts. The burden is on the evaluator to analyze how the components of each new technique differ from concepts already in the literature. The need for this standard is illustrated well by packages for accelerated learning, as discussed in Chapter 4.

Evidence needs also to be adduced that supposed cause and effect variables vary together in a systematic manner. Relevant procedures include comparison of performance before and after introduction of the technique, contrasts of experimental and control groups in an experimental design, and calculation of statistical significance. Illusory covariation can occur more easily in nonstatistical studies, which are used often to support the existence of paranormal effects, as discussed in Chapter 9.

Especially demanding is the need for evidence that the performance effect observed is due to the postulated cause and not to some other variable. Ruling out alternative explanations or mechanisms requires intimate knowledge of a research area. Historical findings and critical commentary are needed to identify alternatives, determine their plausibility, and judge how well they have been ruled out in particular sets of experiments. Common threats to the validity of any presumed cause-

effect relation include effects stemming from subject selection, unexpected changes in organizational forces, the spontaneous maturation of subjects, and the sensitizing effects of a pretest measurement on a posttest assessment. Experiments with random assignment of subjects to treatments are preferred, but some of the better quasi-experimental designs are also useful. Another class of threats to validity is associated with subject reactions to such conceptual irrelevancies as experimenter expectations about how subjects should perform or subjects' performing better merely because they are receiving attention. Procedures that have evolved to reduce this sort of threat include double-blind experiments, placebo control groups, mechanical delivery of treatments, and the elimination of all communication between experimenters and subjects or among subjects. These safeguards, however, are not certain, and implementing them is not a simple matter.

Finally, for a technique to be of value, one must ascertain that a causal relation observed in one setting is likely to be observed in other settings in which the technique is to be employed. Replication of an experiment by an independent investigator is a first step. Another step is to produce the cause and effect with different samples of people, settings, and times. Systematic reviews of the literature, perhaps aided by what is referred to as meta-analysis of studies (as illustrated in Chapter 5), are also helpful. Beyond these steps, a thorough theoretical understanding of causal processes, which is a fundamental goal of science, permits increased practical control.

Our point—perhaps seeming obvious to many but nonetheless needing emphasis here—is that a planned or existing program for implementing an enhancement technique is much more likely to bear fruit if evidence for the technique's effectiveness is properly derived from basic research. A complex set of ground rules exists for conducting and drawing inferences from basic research, and waiving those rules greatly increases the chances of incorrect conclusions.

STANDARDS FOR EVALUATING FIELD TESTS OF PROGRAMS

An adequate appraisal of an actual enhancement program requires attention to three general factors. First, the organizational (i.e., political, administrative) context in which the program is embedded should be described. That context strongly influences the choice of evaluation criteria, the types of evaluations considered feasible, and the extent to which evaluation results will be used. Second, the program's consequences should be described and explained, including planned and unplanned, short-term and long-term consequences. The way the program

is construed influences the claims resulting from an evaluation and the degree of confidence that can be placed in what was learned. Third, value or merit should be explicitly assigned to a program. Valuing relates an enhancement technique to an Army need and to feasible alternatives. In the following sections we comment on these three factors in turn.

The Organizational Context

A description of the broader context of an enhancement program would include an assessment both of the various constituencies with a stake in its implementation and of the priorities of the larger institution. We do not discuss stakeholder interests in general at this point because we refer to some specifically later in this chapter, in the section on the committee's impressions of current Army evaluation practices. We do comment here on the Army's institutional priorities as they may relate to scientific standards.

We understand that the Army, like other organizations in society, may have—and quite possibly should have—different standards for evaluating knowledge claims, or technique effectiveness, than science has. The scientific establishment is conservative in the tests it administers to discipline its conjectures; in particular, its goal is to reduce uncertainty as far as possible, no matter how long that takes. In the Army, by contrast, the need for timely information and decisions may lead to an acceptance of greater uncertainty and a higher risk of being wrong.

There is no Army doctrine of which we are aware concerning the degree of risk that is acceptable in evaluations of pilot programs. Yet surely one objective of evaluations of pilot programs should be to describe the costs to the Army of drawing incorrect conclusions so that inferential standards can be made commensurate with those costs. If the costs are relatively low, the riskier approach of most commercial research (as, for example, in management consulting or marketing) may be preferred to the more conservative approach of basic science.

Describing a Program's Consequences

In evaluating a program, it is desirable to present an analysis and defense of the questions probed and not probed, together with justification for the priorities accorded to various issues. Primary issues usually include the program's immediate effects and its organizational side effects.

Immediate Effects

A primary problem in evaluation is to decide on the criteria by which a program is to be assessed. The major sources for identifying potential

criteria include program goals, interviews with interested persons, consideration of plausible consequences found in the literature, and insights gained from preliminary field work.

Such criteria specify only potential effects, however. They do not speak to the matter of whether the relation between a supposed cause and effect is truly causal. In this respect, a fundamental issue of methodology is the use of randomized experiments. Although logistic reasons abound in any practical context for not going to the trouble to use such research designs, one might nonetheless argue that the Army is in a better position to conduct randomized experiments than are organizations in such fields as education, job training, and public health. The reason for going to such trouble is that randomized experiments give a lower risk of incorrect causal conclusions than the alternatives.

Alternatives at the next level of confidence are quasi-experimental designs that include pretest measures and comparison (control) groups. Relatively little confidence can be placed either in before-after measurements of a single group exposed to a technique without an external comparison, or in comparisons of nonequivalent intact groups for which pretest measures are not available.

Side Effects

Unintended side effects include impacts on the broader organization, and these should be monitored. For example, trainers from other (nonexperimental) units may copy what they think is going on, or they may simply be upset by the implementation of new instructional packages in the experimental units. Units not treated in the same way as the experimental units may be unwilling to cooperate when cooperation would seem to be in their best interest. They may also suffer by comparison, as is thought to be the case, for example, when COHORT units are introduced into a division (see Chapter 8). Evaluators should strive to see any program as fitting into a wider system of Army activities in which it may have unintended positive or negative effects.

Assigning Value to Pilot Programs

The described consequences of a program tell us what a program has achieved but not how valuable it is. Three other factors are important in inferring value: Does the new technique meet a demonstrable Army need to the extent that without it the organization would be less effective? How likely is it that the program can be transferred to other Army settings, either as a total package or in part? How well does the new

program fare when compared with current practi
for bringing about the same results?

Meeting Needs

Representatives of the commercial world wh
products often confound wants with needs, enth
hope with reality. While it is axiomatic that all f
meet genuine Army needs, it is not clea how r
when the developers of new products approac
permission to do general research or fiel tests.
analysis should be part of the documentation abo

What should a needs analysis look lik? At t
document the current level of performance at sc
is inadequate, what reason there is to believe
change, and what the Armywide impacts wou
performance in question were improved. In addi
question why a particular program is needed fo
Such an analysis would describe the program,
justification in basic research, identify the financi
required to make the program work, relate the re
funds available, examine other ways of bringing a
results, and justify the program at hand in terms
effectiveness. To facilitate critical feedback, su
independent of the persons who sponsor progr
thorough, firsthand acquaintance with the progr
and sponsors.

As just described, needs analysis is a plann
mounting a pilot program. It is not a review of
relative to needs, for which a description of a pr
is required. At that later stage in evaluation a judg
whether the magnitude of a program's effects is su
to a degree that makes a practical difference.
whether the program makes a statistically reliable
ance. Size of effect relative to need is the cruci
magnitude of change required for practical significa
in advance, it is easy to use such a specification
need has been met. But the level of change requi
not usually predetermined, and there are political r
are not always eager to have their programs evalu
sizes they themselves have clearly promised or th
them.

Needs can be specified only by Army officials, a

officials inspect the results a program has achieved, relating them to their perception of need. Since the Army is heterogeneous, it would be naive to believe that there are no significant differences within it about how important various needs are and how far a particular effect goes in meeting a particular need. Some theorists relate needs primarily to the number of persons performing below a desired level, while others emphasize the seriousness of consequences for unit performance, for which deficiencies in only one or two persons may be crucial. Some practitioners are likely to think a deficit in skill X is worse than a deficit in skill Y, while others may believe the opposite. Evaluators who take the concept of need seriously have to take cognizance of such heterogeneity, perhaps using group approaches like the Delphi technique to bring about consensus on both the level of need and the extent to which a particular pattern of evaluative results helps meet that need.

Likelihood of Transfer

Although some local commanders may sponsor field trials for the benefit of their command alone, the more widely a successful new practice can be implemented within the Army, the more important it is likely to be. Consequently, evaluations of pilot programs should seek to draw conclusions about the likelihood that findings will transfer to populations and settings different from those studied.

In this regard, it is particularly important to probe the extent to which any findings from a pilot study might depend on the special knowledge and enthusiasm of those persons who deliver or sponsor the program. Such persons are often strongly committed to a program, treating it with a concern and intensity that most regular Army personnel could not be expected to match. While it is sometimes possible to transfer such committed persons from one Army site to another in order to implement a program, in many instances this cannot be done. Transfer is partly a question of the psychology of ownership; authorities who did not sponsor a product will sometimes reject out of hand what others have developed, including their immediate predecessors. Since Army leaders in any position turn over with some regularity due to transfers, promotions, and retirement, successors will probably not identify with a program as strongly as the original sponsors and developers did.

The likelihood of transfer also affects the degree to which program implementation is monitored. Pilot programs are likely to be more obtrusively monitored than other programs. Not only is this obtrusiveness due to developers' and evaluators' fussing over their charge, it is also due to teams of experts brought in to inspect what is novel and to responsible officers wanting to show others the unique programs they

are leading (and on which the success of their careers may depend). For at least these reasons pilot programs tend to stand out more than the regular programs they may engender. Research suggests that the quality with which programs are delivered may in fact increase when outside personnel are obviously monitoring individual and group performance.

It is naive to believe that one can go confidently from a single pilot program to full-blown Armywide implementation. Even if this were feasible politically, it would not be technically advisable unless there were compelling evidence from a great deal of prior research indicating that the program was indeed built on valid substantive foundations. Given a single pilot program, decisions about transfer are best made if the program is tested again, at a larger but still restricted set of sites and under conditions that more closely approximate those that would pertain if the new enhancement technique were implemented as routine policy. Only then might serious plans for Armywide implementation be feasible.

Contrast with Alternatives

Most of the evaluation we have discussed contrasts a novel program with standard practices that are believed worth improving; yet rational models of decision making are usually predicated on managers' having to choose among several different options for performing a particular task. One would hope that every sponsor of a novel performance enhancement technique is conversant with the practical alternatives to it and has cogent arguments for rejecting them.

Many novel techniques have some components that are already in standard practice or can be clearly derived from established theories. Upon close inspection, pilot programs often turn out to be less novel than their developers and sponsors claim. Of course, the Army may often find it convenient to order complete packages in the form offered and may not have much latitude to interact with developers in order to modify package contents to emphasize what is truly a novel alternative and to downplay that which is merely standard practice.

Ultimately, alternatives have to do with costs. Although many forms of cost are at issue—including those associated with how much a new practice disrupts normal Army activities and how much stress it puts on personnel—the major cost usually considered is financial. Cost analysis is always difficult, nowhere more so than in the Army, which uses many ways to calculate personnel costs. Nonetheless, in planning an evaluation, some evidence about the total cost of a pilot program to the Army will usually be available and can be critically scrutinized. It is also useful, as far as possible, to ascribe accurate Army costs to each of the major components of such an intervention. In our view, what is called cost-

effectiveness analysis lends itself better than what is called cost-benefit analysis to the comparison of different programs. The purpose of cost-effectiveness research is to express the total cost for each program in dollar terms and to relate this to the amount of effect as expressed in its original metrics—unlike cost-benefit research, in which even the effects have to be expressed in dollar terms. Sophisticated consumers of evaluation should want something akin to cost-effectiveness knowledge, for it reflects decisions they should be making. Is it not useful to know, for example, that the best available computer-assisted instruction packages are much less cost-effective than peer tutoring?

CURRENT STATUS OF ARMY EVALUATIONS

We set forth here some of our impressions of the way in which the Army currently manages the solicitation and evaluation of novel techniques to enhance performance. We must stress that these are only impressions, gained through the limited investigative capabilities of a committee such as ours, not hard conclusions based on systematic research directed at the particular question. Furthermore, although the opinions that follow are largely critical of Army procedures, they are not accompanied by much detail. As noted earlier, the focus here is on the identification of the various Army constituencies that have a stake in enhancement programs and on the role they play in evaluation.

How the Army decides which among competing proposals should be sponsored for development or for field tests is not clear. What is clear is that decision making is diffuse both geographically and institutionally. Sponsorship may come from senior managers in the Pentagon or from local personnel of varying rank. While differences in the quality of program design, implementation, or evaluation may be correlated with the source of sponsorship, such a correlation is not clear at present in the Army context.

A particular concern is that Army sponsors of pilot programs may base their judgment about the value of a program either on their own ideas about what is desirable or effective or on the persuasiveness of the arguments presented to them by program developers, who stand to gain financially if the Army adopts their program. Judgments of value should depend on broader analysis of Army needs and resources, as well as on realistic assessment of the quality of proposed ideas based on a thorough and independent knowledge of the relevant research literatures. Sponsors should examine what is being advocated at every stage: proposal, testing, and implementation.

Also of concern when pilot programs are planned is how decisions are reached about funding and about the quality of implementation expected

from them. Although systematic evidence is lacking, it seemed to committee members that pilot programs are not generally implemented well and, except for fiscal accountability, are not closely monitored by their Army sponsors. Evaluations of pilot programs should try to characterize resources required by the program and the resources actually available.

We found little evidence that sponsors, advocates, or local implementers had aspirations to evaluations that use state-of-the-art methods. We found no guidelines about the standards expected for evaluative work, whether in the form of published minimal standards or published statements of preferred practices. When it comes to field trials of novel ideas for enhancing human performance, the monitoring of evaluation quality does not seem to be part of the organizational context. Given the absence of formal expectations in these regards, it is not surprising that the pilot programs we saw and the evaluation materials we read were usually disappointing in the technical quality of the research conducted. In settings in which program sponsors or advocates control an evaluation, weaker evaluations (e.g., based on testimony) will sometimes be preferred to stronger methods (e.g., experiments) because the latter are usually more disruptive when implemented and are more likely to result in effects that are disappointing, however much more accurate they may be. The weaker methods are easier to implement when few units are available, are less disruptive of ongoing activities, are easier to manipulate for self-interested ends, and need not be as expensive for data collection.

We saw little evidence that the Army requires evaluations by persons independent of the pilot program under review. Moreover, the nonindependent evaluations we saw did not seem to have been subjected to any of the peer review procedures to which research results (and plans) are subjected not only in academic sciences, but also in much of the corporate world, as with, say, pharmaceutical testing. While in-house evaluation is highly valuable for gaining feedback for program improvement, many experienced evaluators contend that it is inadequate for assigning overall value because in-house evaluators cannot divorce themselves from their own stake in the program under examination. Although it is not easy to specify organizational standards adequate for a high-quality field test of some novel technique, it is also not difficult to detect the inadequacies associated with local program sponsors' having few clear expectations about the desirable qualities of program operations or evaluative practices. In the absence of such expectations, program developers and evaluators may believe that few officials care about the small-scale field tests of techniques on which the developers'—and, all too often, the evaluators'—own welfare depends.

Since the organizational climate we have just described is not optimal

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

for gaining trustworthy information about program value, future evaluators of Army field trials might do well to characterize: (1) what program managers expect in terms of the quality of the program and its evaluation; (2) who is paying attention to the trials; and (3) for what purposes they want to use any information provided by the evaluation. This kind of information, as mentioned above, contributes to a description of the organizational context of a program, which is a major part of an adequate evaluation.

QUALITATIVE APPROACHES

Alternatives to experimentation are the largely qualitative traditions, which rely mostly on direct observation, sometimes supplemented by archival data. Investigative journalists operate in this mode; so do many cultural anthropologists, political scientists, and historians. These professions use clues to suggest hypotheses about possible causes and investigate the empirical evidence in ever-greater detail in an attempt to rule out hypotheses until they are left with just one. A critical aspect of their work is the use of substantive theories and ad hoc findings from the past to help in ruling out alternative explanations. Also working in this tradition are committees of psychologists who seek to make statements about the causes of enhanced human performance. Rarely conducting studies themselves, they instead sift through historical evidence provided by reviews of the literature and make on-site observations in the manner of detectives, pathologists, investigative journalists, and cultural anthropologists.

These traditions rely strongly on personal testimony. Respondents' reports are taken seriously and, indeed, should be. Any method can, in principle, generate strong causal evidence, provided that plausible alternatives to a preferred hypothesis have been ruled out. The general issues are: Can personal testimony usually rule out all the plausible alternative interpretations? Does use of it engender the very threats to validity that militate against strong inferences? Dale Griffin, in a paper prepared for the committee (see Appendix B), suggests "no" to the first question and "yes" to the second. His analysis of biases that operate when people attempt to explain how and why they changed after an experience reveals many of the shortcomings associated with relying on testimony as a major means of testing causal hypotheses.

While testimony can be regarded as a form of confirmatory evidence, it does not provide any of the disconfirming evidence needed to reduce uncertainty. Rarely are there the kinds of comprehensive probes needed to discover why respondents believe that the effects are due to a treatment rather than to maturation, statistical regression, or the pleasant feelings

aroused by the experiences. People are typically weak at identifying the range of such alternatives, however simply they may be described, and at distinguishing the different ways in which the causal forces might operate. How can people know how they would have matured over time in the absence of an intervention (technique) that is being assessed? How can people disentangle effects due to a pleasant experience, a dynamic leader, or a sense of doing something important from effects due to the critical components of the treatment per se? Much research has shown that individuals are poor intuitive scientists and that they recreate a set of known cognitive biases (Nisbett and Ross, 1980; Griffin). These include belief perseverance, selective memory, errors of attribution, and overconfidence. These biases influence experts and nonexperts alike, usually without one's awareness of them. Scientists hold these biases in partial check by using random assignment instead of testimony and by the tradition of public scrutiny to identify and analyze alternative interpretations for observed events. Such methodological traditions can be transmitted to consumers and producers of enhancement techniques through courses on statistical inference and formal decision making. These courses would have the salutary effect of calling attention to the shortcomings of testimony as evidence.

We submit that experimental methods facilitate causal inferences better than the alternatives. They reduce more uncertainty by ruling out more of the contending interpretations for observed effects. However, we refer here to the *relative* superiority of experimentation; such superiority should not be confused with either the perfection or even the adequacy of experimentation. Its problems include the facts that experiments cannot be implemented under all conditions and that experimentation has its own set of unintended side effects. Thus, experimental methods do not guarantee causal inferences and so cannot obviate the need for critical analysis that, on a case-by-case basis, is sensitive to the contexts and traditions of particular institutions or communities, such as the Army, on one hand, and the various promoters of new enhancement techniques, on the other. Moreover, well-conceived research is costly: it requires specially trained investigators, equipped facilities, and programs that may need extensive collaborations and review panels. It is also a demanding craft that requires sensitivity to detail and precision in order to ensure results that are interpretable.

On balance, the benefits derived from careful experimentation outweigh the costs just mentioned. All other things being equal, experimentation is much the preferred strategy for judging the efficacy of techniques that purport to enhance performance, and it should be used whenever possible.

PART III

Parapsychological Techniques

OF ALL THE SUBJECTS TREATED in this volume, none is more controversial than parapsychology. While the flavor of the debates is captured to some extent in this chapter, the subject is treated in the same manner as the other techniques reviewed: we address the question of whether the evidence warrants further consideration of parapsychological techniques for research or application or both.

Emphasized here is information gathering by remote viewing and mind-over-matter effects in controlling machine behavior, particularly machines that generate series of random numbers, which are often used in parapsychology experiments. Although scattered results are said to be statistically significant, an evaluation of a large body of the best available evidence does not support the contention that these phenomena exist. If, however, future experiments, conducted according to the best possible methodological standards, are more generally viewed as producing significant results, it would be appropriate to consider a systematic program of research. Such a program should include a concern for the need to proceed from small effects to practical applications.

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

9

Paranormal Phenomena

BACKGROUND

The primary purpose of this chapter is to evaluate the scientific evidence on parapsychological techniques in selected areas. A more complete understanding of the topic, however, requires that we provide background on the military's interest in these phenomena and treat the conceptual issue of how people come to believe as they do. This background section includes a discussion of the phenomena and the military's interest in them as well as an overview of the committee's focus. A brief examination of the different kinds of justifications for the claims is followed by a more detailed treatment of the evidence in areas that have produced large literatures: remote viewing, random number generators, and what are called Ganzfeld (whole visual field) experiments. In addition, we describe experimental work that the committee actually witnessed by visiting a parapsychological laboratory. Despite the growing scientific tradition in some of these areas, many people continue to rely on qualitative or experiential evidence to support their beliefs; we discuss the problems associated with qualitative evidence in conjunction with the research on cognitive and emotional biases, which is reviewed in the paper by Dale Griffin (Appendix B). Finally, the chapter summarizes the committee's major conclusions.

The Nature of the Phenomena

Parapsychologists divide *psi*—the term applied to all psychic phenomena—into two broad categories: *extrasensory perception* (ESP) and

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

psychokinesis (PK). Included in ESP are telepathy, precognition, and clairvoyance, all of which refer to methods of gathering information about objects or thoughts without the intervention of known sensory mechanisms. Popularly called mind over matter, PK refers to the influence of thoughts upon objects without the intervention of known physical processes.

A presentation to the committee by several military officers described in some detail the results of experiments in remote viewing carried out at both SRI International and the Engineering Anomalies Research Laboratory at Princeton University. In these experiments subjects are said to have more or less accurately described a geographical location being visited by a target team. Although the human subjects have no way of normally knowing the target location, the examples recounted appear to indicate, at first glance, some striking correspondences between their descriptions and the actual sites. These studies have been related by some persons to reported out-of-body experiences.

The presentation included discussion of psychic mind-altering techniques, the levitation claims of transcendental meditation groups, psychotronic weapons, psychic metal bending, dowsing, thought photography, and bioenergy transfer. It was indicated that the Soviet Union is far ahead of the United States in developing potential applications of such paranormal phenomena, in particular psychically controlling and influencing minds at a distance. At the presentation, personal accounts were given of spoon-bending parties, in which participants believe they have caused cutlery to bend with the power of their minds, as well as instances of self-hypnosis to control pain and cure illness, walking barefoot on fire and handling hot coals without being burned, leaving one's body at will, and bursting clouds by psychic means.

The media and popular publications, especially in recent years, have discussed various aspects of psychic warfare. Three recent books, by Ebon (1983), McRae (1984), and Targ and Harary (1984), have attempted to document Soviet and American efforts to develop military and intelligence applications of alleged paranormal phenomena. These accounts have been augmented by newspaper stories, magazine articles, and television programs. Many of these sources acknowledge the speculative nature of the proposed applications, but others report that some of the techniques already exist and work.

The claimed phenomena and applications range from the incredible to the outrageously incredible. The "antimissile time warp," for example, is supposed to somehow deflect attack by nuclear warheads so that they will transcend time and explode among the ancient dinosaurs, thereby leaving us unharmed but destroying many dinosaurs (and, presumably, some of our evolutionary ancestors). Other psychotronic weapons, such as the "hyperspatial nuclear howitzer," are claimed to have equally bizarre capabilities. Many of the sources cite the claim that Soviet psychotronic weapons were responsible for the 1976 outbreak of Legionnaires' disease, as well as the 1963 sinking of the nuclear submarine *Thresher*.

Potential Military Applications

Some people, including some military decision makers, can imagine potential military applications of the two broad categories of psychic phenomena. In their view, ESP, if real and controllable, could be used for intelligence gathering and, because it includes "precognition," ESP could also be used to anticipate the actions of an enemy. It is believed that PK, if realizable, might be used to jam enemy computers, prematurely trigger nuclear weapons, and incapacitate weapons and vehicles. More specific applications envisioned involve behavior modification; inducing sickness, disorientation, or even death in a distant enemy; communicating with submarines; planting thoughts in individuals without their knowledge; hypnotizing individuals at a distance; psychotronic weapons of various kinds; psychic shields to protect sensitive information or military installations; and the like. One suggested application is a conception of the "First Earth Battalion," made up of "warrior monks," who will have mastered almost all the techniques under consideration by the committee, including the use of ESP, leaving their bodies at will, levitating, psychic healing, and walking through walls.

The Committee's Focus

Although such colorful examples provide the context for our agenda, the cumulative body of data in the discipline of parapsychology enables us to judge the degree to which paranormal claims should be taken seriously. Since 1882 reports of both naturally occurring incidents and phenomena in laboratory settings have been accumulated in journals, monographs, and books. Just to survey the reports in the refereed journals of parapsychology would be an enormous undertaking. As scientists, our inclination is, of course, to restrict ourselves to the evidence that purports to be scientific. But the alleged phenomena that have apparently gained most attention and that have apparently convinced many proponents do not come from the parapsychological laboratory. Nothing approaching a scientific literature supports the claims for psychotronic weaponry, psychic metal bending, out-of-body experiences, and other potential applications supported by many proponents.

The phenomena are real and important in the minds of proponents, so

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

we attempt to evaluate them fairly. Although we cannot rely solely on a scientific data base to evaluate the claims, their credibility ultimately must stand or fall on the basis of data from scientific research that is subject to adequate control and is potentially replicable.

We divided the task into two parts. First, we looked at the best scientific arguments for the reality of psychic phenomena. Our sponsors, as well as our own appraisal of the current status of parapsychology, indicated that the two most influential scientific programs were the experiments on remote viewing and the experiments on psychokinesis using random event generators. In addition, we looked at the research on the Ganzfeld (whole visual field) because this, in the opinion of many parapsychologists, is the most likely candidate for a replicable experiment. We also report on a parapsychological experiment that the committee itself witnessed.

Second, we considered the arguments of proponents who rely on what they call qualitative as opposed to quantitative evidence for the paranormal. Such evidence depends on personal experience or the testimony of others who have had such experience. Most, if not all, of this evidence cannot be evaluated by scientific standards, yet it has created compelling beliefs among many who have encountered it. Witnessing or having an anomalous experience can be more powerful than large accumulations of quantitative, scientific data as a method of creating and reinforcing beliefs. Because personal experience rather than scientific data has been the source of most beliefs in the paranormal, we have devoted some of our resources to considering this sort of cognitive method as a tool for achieving knowledge.

STANDARDS OF EVIDENCE

Diverse justifications have been offered for pursuing paranormal claims. One argument asserts that paranormal phenomena may no longer be anomalous, given the implications of contemporary quantum mechanics. Indeed, a few physicists have supported some parapsychologists in maintaining that certain forms of precognition and psychokinesis are consistent with some interpretations of quantum theory. The other major argument is that we have no choice but to get involved because the Soviet Union already has a program to develop military applications of psychic phenomena.

Several proponents, including some scientists, firmly believe that paranormal phenomena have been scientifically demonstrated several times over. At the same time, most scientists do not believe that psi exists. Many persons on both sides believe this paradox to be the result of irrational and dogmatic belief systems. The proponents accuse the critics of being closed-minded and bigoted. The critics imply that the

proponents have allowed wishful thinking to bias their judgment and that they are incompetent scientists and are self-deceived. Both sides can point to examples to back their positions.

One essential question confronts the committee: What does an impartial examination of the scientific evidence reveal about the existence of psi? Such an examination assumes that clear standards exist for judging the adequacy of the evidence, which, in turn, raises the issue of what constitutes sufficient evidence. That issue involves many difficult philosophical, theoretical, and methodological matters. For example, Palmer, in his "An Evaluative Report on the Current Status of Parapsychology" (1985), denies that current parapsychological experiments can provide any evidence for the existence of psi. This is because psi implies paranormality and, according to Palmer, we cannot argue that a given effect has a paranormal cause until we have an adequate theory of paranormality. He further argues, however, that parapsychological experiments can and do provide evidence for the existence of anomalies. By an anomaly, Palmer means a statistically significant deviation from chance expectation that cannot readily be explained by existing scientific theories. The burden of Palmer's paper is that just such anomalies have been demonstrated.

Because parapsychologists other than Palmer do not make this distinction between demonstrating an anomaly and testing a theory of paranormality, we do not carry on this distinction in our own assessment of the evidence. We tend to agree with Palmer on this matter, however. When we talk about evidence for psi in the remainder of this chapter, we are using psi in the neutral sense of an apparent anomaly rather than in the stronger sense of a paranormal phenomenon.

Minimal Criteria

Fortunately, critics and parapsychologists appear to agree on the general requirements necessary to demonstrate psi in a parapsychological experiment. Both Palmer (1985) and James E. Alcock (Appendix B) discuss such criteria in their respective papers. As Palmer points out, psi is defined negatively as a statistical departure from a chance baseline that cannot be accounted for by chance, sensory cues, or known artifacts. Such a negative definition implies the minimal criteria required to justify a conclusion that psi has been demonstrated.

Given the statistical aspect, it is imperative that the data be collected in such a way that the underlying probability model and assumptions of the statistical test are fulfilled. This means that targets must be adequately randomized and that each trial in the experiment must be independent of the preceding ones—and, of course, the statistical procedures must be

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

applied and interpreted correctly. Given that all ordinary explanations must be ruled out, the experimenter must take special precautions to ensure that sensory cues, recording errors, subject fraud, and other alternatives have been prevented. Although it is impossible to rule out completely every possible contaminant or to anticipate every alternative, there are reasonable standards that most parapsychologists would agree should be followed.

Because different research paradigms have their own special requirements, no single set of standards can be specified in advance for all parapsychological experiments. Experiments with electronic number generators, for example, rarely have problems with data recording, but they do require special methods such as tests of randomness and attention to the immediate physical environment that are unnecessary with more traditional parapsychological experiments. One requirement for assessing the adequacy of a given experiment is that its procedures and methods of analysis be adequately documented. Unless we know how the targets were selected, how the results were analyzed, how the possibility of sensory leakage was prevented, and how other such aspects of the study were carried out, we have no basis for evaluating the quality of the information provided by the experiment.

Global Criteria

The criteria mentioned in the preceding paragraphs apply to the individual experiment. More global criteria come into play when one wants to evaluate an entire research program or set of experiments. Here we look for such things as replicability, robustness, lawfulness, manipulability, and coherent theory. These criteria deal with the coherence and intelligibility of the alleged phenomena. It is in terms of such global criteria that parapsychological research has been especially vulnerable. Much of the objectivity involved in assessing the adequacy of research applies to judging individual experiments. But science is cumulative and depends not so much on the outcome of a single experiment as on consistent and lawful patterns of results across many experiments carried out in a variety of independent settings. Lawful consistency in this sense, according to both parapsychologists and their critics, has never been found in parapsychological investigations in the history of psychic research. Recently a few parapsychologists have expressed the hope that the experiments on remote viewing, random number generators, and the Ganzfeld (the very ones we have chosen to examine in detail in this report) may actually yield the long-sought replicability. The type of replicability that has been claimed so far is the possibility of obtaining significant departures from the chance baseline in only a proportion of

the experiments, which is a kind of replicability quite different from the consistent and lawful patterns of covariation found in other areas of inquiry.

Despite the fact that scientific progress in a given area depends on the accumulation of lawful and consistent patterns across many experiments, the methods for deciding that such consistency exists are still quite primitive in comparison with the standards for judging the adequacy of a single experiment. Indeed, it is only within the past few years that serious attention has been devoted to developing objective and standardized procedures for evaluating the consistencies across a body of independent studies. For the most part, judgment about what a body of investigations demonstrates is still a surprisingly intuitive and haphazard process. This probably has not been a serious drawback in those areas of inquiry in which the basic phenomena are robust and experiments can be conducted with high confidence that the predicted relations will be obtained; but such impressionistic means for aggregating the outcomes of several experiments in the domain of parapsychology open the door to all the motivational and cognitive biases discussed in the paper prepared for the committee by Griffin. Not only are the data and alleged correlations erratic and elusive in this field, but their very existence is open to question.

EVALUATION OF THE SCIENTIFIC EVIDENCE

To evaluate the best scientific evidence on the existence of psi, and with the advice of proponents and our sponsors, we conducted site visits to some of the most notable parapsychological laboratories. The parapsychology subcommittee (see Appendix C) visited Robert Jahn's Engineering Anomalies Research Laboratory at Princeton University, where it witnessed presentations and demonstrations regarding psychokinetic experiments on random number generators. Jahn and his associates also briefed the subcommittee on the current status of their work in remote viewing.

The subcommittee also visited Helmut Schmidt's laboratory at the Mind Science Foundation, San Antonio, Texas. Schmidt pioneered the use of random number generators in parapsychology experiments in 1969. His is considered one of the two major research programs on psychokinesis (the second is Jahn's).

As an additional posssible input, the committee agreed to participate in a psychokinetic experiment of new design with Helmut Schmidt. Specifically, Schmidt accepted the suggestion that the committee's consultant, Paul Horwitz, be included in the conduct of the experiment. The

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

work has not yet begun, however, and it now appears that we will not have any results to report before our terms expire.

The chair of the parapsychology subcommittee also visited SRI International, another major laboratory studying psychic effects on random number generators. (This latter research group argues that the observed effects are not due to psychokinesis but rather represent a special form of precognition.) The subcommittee chair also attended the meetings of the Parapsychological Association held at Sonoma State College in California. The entire committee made a site visit to Cleve Backster's laboratory in San Diego (arranged to coincide with the committee's meeting in La Jolla, California).

These site visits enabled the committee to observe firsthand the experimental arrangements and equipment used by some of the major contributors to parapsychological research. They also provided us an opportunity to discuss results, interpretations, and problems with a few important investigators. We were impressed with the sincerity and dedication of these investigators and believe that they are trying to conduct their research in the best scientific tradition. We also got the impression that this type of research involves many unresolved problems and still has a long way to go before it develops standardized, easily replicable procedures. The information obtained from these site visits does not provide an adequate basis for making scientific judgments. For this we rely, as we would in other fields of science, on a careful survey of the literature.

Research on Remote Viewing

The SRI Remote Viewing Program

Since the early 1970s, probably the best known research program in parapsychology has been the experiments in remote viewing initiated by physicists Harold Puthoff and Russell Targ when they were at SRI International. In a typical remote viewing experiment a subject, or percipient, remains in a room or laboratory with an experimenter, while a target team visits a randomly selected geographical site (e.g., a shopping mall, an outdoor arena, the Palo Alto airport, the Hoover tower). Neither the experimenter nor the subject has been given any information about the target. Once the experimenter and the subject are closeted in the laboratory, they wait for 30 minutes before the subject begins to describe his or her impressions of the target site.

Meanwhile the target team, consisting of two to four members of the SRI staff, obtains instructions for going to a randomly chosen target site from another SRI staff member. They then drive to the designated target site and remain there for an agreed-on 15-minute period (after allowing approximately 30 minutes to reach the site). During the time that the target team remains at the target site, the subject describes his or her impressions into a tape recorder and also makes any drawings that would help to clarify those impressions. When the target team returns to the laboratory, all the participants listen to the tape recording of the subject's impressions. Then all the participants go to the target site, where the subject is allowed to see how closely his or her impressions agreed with the actual target.

The first subject to participate in such a formal series of trials was the late Pat Price. In the first series, consisting of nine sessions, the duration of each session was 30 minutes. The transcript for each session is rich in detail; the one published transcript in Targ and Puthoff's first book runs to almost six printed pages (Targ and Puthoff, 1977).

Given such data, how does one decide if the experiment was a success? Did Price's descriptions, for example, convey correct knowledge of the different target sites? In fact, two methods have been used to demonstrate the effectiveness of remote viewing. One method is simply to compare the description with the target and make a judgment as to whether the correspondence is sufficient to claim a "hit." The second method uses an independent judge to rank the degree to which each description matches each site and then applies statistical tests to decide if the association is greater than chance.

Unprecedented success was claimed for the early remote viewing experiments in terms of both methods (Targ and Puthoff, 1974, 1977; Puthoff and Targ, 1976). Many examples were supplied of dramatic correspondences between impressions of the percipient and the physical details of the actual target. Such correspondences, no matter how dramatic and compelling, do not carry scientific weight, because it is impossible to assess their probabilities. In addition, much psychological research indicates how such subjective validation can create strong, but false, illusions of matching (see below).

The more formal evidence from the rankings of independent judges was also impressive. The first formal series of nine trials resulted in seven of the transcripts being ranked 1 against their intended target sites by the independent judge. Only one such ranking would be expected by chance. Puthoff and Targ reported the probability of such an outcome being due to chance as only 0.0000029. The second formal series, using Hella Hammid, was equally impressive, producing five first places and four second places in the rankings of transcripts against target sites.

Although subsequent series by Targ and Puthoff, as well as by

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

other investigators, have not always yielded such overwhelmingly impressive results, most of them have continued to display highly significant outcomes (Targ and Harary, 1984). On the surface, at least, this is a reliable, simple, and highly effective recipe for producing paranormal communication. Especially appealing is the claim that remote viewing works with just about everyone. Targ and Harary, for example, provide exercises for anyone who wants to develop and improve his or her ability to pick up information at remote sites. Neither space nor time, its proponents assert, is a barrier. The percipient can pick up information from the surface of Jupiter as well as from target sites that can be visited at some future time.

Scientific Assessment of Remote Viewing

After the first remote viewing experiments were conducted in the early 1970s, many investigators throughout the world tried to follow suit. Most of them believed that their findings supported the claims of the SRI International researchers. The majority of these experiments, however, consisted of informal demonstrations rather than formal scientific experiments and relied solely on subjective matching. In the past 15 years, the number of formal experimental replications of the SRI remote viewing experiments has been surprisingly few.

Targ and Harary (1984) include as an appendix in their book a report by Hansen, Schlitz, and Tart that evaluates all the known remote viewing experiments conducted from 1973 through 1982. "In an examination of the twenty-eight formal published reports of attempted replications of remote viewing," write Targ and Harary, "Hansen, Schlitz, and Tart at the Institute for Parapsychology found that more than half of the papers reported successful outcomes." They concluded: "We have found that more than half (fifteen out of twenty-eight) of the published formal experiments have been successful, where only one in twenty would be expected by chance."

Two comments may be in order with respect to the foregoing conclusion. First, given the enormous publicity and the unusually strong claims, 28 formal experiments in 10 years seems surprisingly few. In comparison, the Ganzfeld psi experiments produced approximately twice as many formal experiments during the same interval. Second, 13 of the 28 formal experiments, or 46 percent, failed to claim successful outcomes. This rate of failure is much higher than what might have been expected on the basis of the earlier claims by Targ and Puthoff (1977), namely, that they had succeeded with every subject they had tried.

Even 15 successful outcomes out of 28 tries is impressive, especially by parapsychological standards. An inspection of the listed studies, however, suggests that the 28 formal experiments vary considerably in their importance. Some of these "published formal experiments" appeared as brief reports or abstracts of papers delivered at meetings of the Parapsychological Association or similar organizations. Others appeared in print only as brief or informal reports in book chapters or letters to the editor. Altogether, 15 of the 28 were published under conditions that fall short of scientific acceptability. Only 13, or 46 percent, of the experiments were published under refereed auspices. As in other sciences, only published reports that have undergone peer review and are adequately documented can be considered seriously as part of the scientific data base.

Of the 13 scientifically reported experiments, 9 are classified as successful in their outcomes by Hansen et al. (Targ and Harary, 1984). Seven of these nine experiments were conducted by Targ and Puthoff at SRI International, the remaining two at other laboratories. This relatively small harvest of nine "successful" experiments suffers from the fact that each is seriously flawed. A variety of problems afflicts the published reports on remote viewing. The documentation, even according to many parapsychologists, is seriously inadequate. Attempts by both neutral and skeptical investigators to gain access to the raw data have typically been thwarted or strongly resisted. Because the essence of scientific justification is public accessibility to the data, this relative inaccessibility suggests that much of the remote viewing data base is not part of science.

Most of the reasons for questioning the acceptability of the evidence for remote viewing lie in a methodological flaw that characterizes all but one of the experiments deemed successful: the successive trials are not independent of one another. This lack of independence has unfortunate consequences for any attempt to draw conclusions about ESP based on the outcomes of such experiments. The concept of independence is technical and somewhat difficult to explain simply, but, since it is critical to understanding why the remote viewing experiments fail to make their case, we supply an intuitive explanation.

Assume that we are considering a remote viewing experiment in which the subject participates in only two trials. In other words, we deal with two randomly chosen target sites. For the first trial, the target team goes to the first target site and remains there while the subject produces his or her first description. Immediately after this trial, the target team returns to the laboratory and takes the subject to the actual target site so that he or she and the others can gain a

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

subjective impression of how closely the description corresponds with the target. For the second trial, the target team visits a second randomly chosen site. While they are visiting this site, the subject produces a second description.

When the experiment is over, the list of target sites (in random order) and the transcripts of the subject's descriptions are given to a judge, who also visits each site. While at a given site, the judge reads the two transcripts and ranks them in terms of how well each one corresponds with the particular site. In our example, one of the transcripts will be ranked 1 and the other will be ranked 2 (with 1 indicating the better correspondence between that target and the transcript). After visiting one site and doing this ranking, the judge then visits the second site and repeats the ranking procedure. The raw data can be set out in a matrix with the target sites as the columns and the transcripts as the rows.

A perfect outcome would be indicated if the transcript produced at the time the team was visiting site A was ranked 1 against that site, and the transcript produced when the team was visiting site B was ranked 1 for that site. (Of course, two trials would be too few to make an adequate statistical assessment of the success of the matching—successful matching would occur too frequently just by chance. The principles we want to illustrate, however, remain the same for two as for many trials.)

If the successive trials in the experiment were independent of one another, and we were interested only in direct hits (that is, outcomes for which the intended transcript was rated 1 against the target site), then we could expect the subject to make between zero and two direct hits. Indeed, if chance alone were operating, there would be four, equally likely, possibilities: (1) no hits, (2) a hit on the first trial and a miss on the second, (3) a miss on the first trial and a hit on the second, and (4) two hits. By this reckoning, the subject could be expected to get two direct hits just by chance in one of every four experiments.

But, as we indicated, the successive trials are not independent. This is because the judge is almost certainly not going to rank a transcript as 1 for more than one target site. This means, in our example, that if he or she ranks the first transcript 1 for target A, then he or she will probably rank the second transcript 1 for target B. In effect, this lack of independence between trials means that, instead of four equally likely possible outcomes there are only two: no hits or two hits. The dependence between trials has created a situation in which the chance probability of two hits is now 50 percent rather than 25 percent.

In this situation, if an experimenter uses a statistical test that assumes independence, he or she will come out with the wrong probabilities. In fact, the statistical test will exaggerate the significance of many outcomes. The failure of the experimenters to realize this problem resulted in exaggerated levels of significance for the early remote viewing experiments. Kennedy (1979), who originally pointed to this problem, recalculated the probabilities for some of these experiments. Puthoff and Targ (1976) reported that five of their first six remote viewing experiments were significant at the .05 level. With Kennedy's corrections for lack of independence, only two remained significant. According to Kennedy, only one of the two successful replications by Bisaha and Dunne (1979) remained significant with the more appropriate test.

One reason for the optimistic initial beliefs in the scientific reality of remote viewing was the fact that the lack of independence between trials produced exaggerated odds against chance results. But even with conservative corrections for lack of independence, approximately one-third of the early experiments still yielded successful outcomes.

One easy way to avoid this problem of dependence is to use a separate target pool of possible sites for each trial. For example, for the first trial one could designate a pool of four possible sites, one of which is randomly chosen to be the actual target site. A second pool of four different possible sites would be used for the second trial. When the trials are completed, the judge is given the list of the four sites for the first trial along with the subject's description for that trial. The judge then ranks each site in terms of its correspondence to the description. The four possible sites for the second trial are then ranked in terms of their correspondence to the subject's description for the second trial. In this illustration, the subject has a probability of 1 in 4 of having the actual target site ranked 1 on each trial, or a probability of 1 in 16 of being correct on both trials.

This second procedure, which is typically used in most free-response parapsychological experiments (such as the Ganzfeld experiments discussed below), not only guarantees independence between successive trials, but also avoids other serious problems, which we discuss next. The fact that the subject is given feedback by being taken to the target site immediately after each trial creates an additional form of dependence between trials. For this reason, other possibilities exist for obtaining "successful" results artifactually. The transcripts can contain clues that provide nonparanormal reasons for judges to associate descriptions with targets correctly. Some of these

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

clues can be quite overt, such as when a subject mentions in the description how the current target apparently differs from a previous target site. When such a clue appears in the description, it provides the judge with information that the current description does not belong with the previous site. This increases the probability that the description will be matched with its appropriate target.

Marks and Kammann (1978) initiated a controversy, still not fully resolved, by claiming that such overt clues were sufficient to account for the striking results of the very first SRI remote viewing with Pat Price. Targ and Puthoff did not deny the existence of such clues in the Price series but argued that they were not sufficient to have accounted for the results. This dispute still has not been settled (Targ, Puthoff, and Targ, 1980; Scott, 1982; Marks and Scott, 1986).

Possibly this controversy over the role of the more overt clues has deflected attention from a much more fundamental and fatally damaging criticism first made by Hyman (1979) and independently by Kennedy (1979). Hyman and Kennedy pointed out that the combination of immediate feedback and lack of independence between successive trials makes it virtually impossible to prevent sensory cueing in the transcripts. As long as both the subject and the experimenter who is closeted with the subject are not blind to the preceding target sites, there is no way to prevent the transcript from being affected in a variety of possible and perhaps subtle ways by the knowledge of the preceding targets.

Hyman (1984–1985) provides an illustration of how such implicit sensory cueing might occur (pp. 131–132):

> Say that the target for the first session was the Hoover Tower at Stanford. This will almost certainly influence what both the viewer and the interviewer say during the second and subsequent sessions in the same series. Almost certainly the viewer, during the second session, will not supply an exact description of the Hoover Tower. So, whatever the viewer says during the second session, a judge should find it to be a closer match to the second target site than to the first one. Now, assume that the second target site happened to be the Palo Alto train station. The viewer's descriptions during the third session will avoid describing either the Hoover Tower or the Palo Alto train station. We do not need to hypothesize something as mysterious as psi to predict that a judge should find this third description a better match to the third target site than either of the first two. As we add sessions, this effect of immediate feedback should continue to make the correlation between the viewer's descriptions and the target sites better and better.

No amount of editing for overt clues can overcome this defect of remote viewing experiments that follow the SRI pattern of dependent trials and immediate feedback. The mechanism described by Hyman

should result in some dramatic correspondences. These dramatic correspondences, in conjunction with subjective validation, are a highly potent recipe for creating the illusion (for both experimenters and subjects) that ESP has occurred.

Palmer (1985), a major parapsychologist who otherwise carefully considers the criticisms of parapsychology, misses the seriousness of this flaw. In mentioning Hyman's criticism, he writes (p. 50):

> It has been suggested by Hyman (1979) that since the subjects in most cases received feedback of the correct target after each trial, the subject could have gained some advantage by avoiding to mention characteristics of targets in earlier trials in their responses in later trials. As noted by Targ, Puthoff, and May (1979), the target pool for the geographical-site experiments was sufficiently large and contained sufficient redundancy that this is unlikely to be a significant biasing factor.

Perhaps such complacency has enabled experimenters to continue conducting remote viewing experiments with this fatal flaw. In fact, the size of the target pool, no matter how large, does not affect the validity of Hyman and Kennedy's criticism. Nor does the claim that the pool contained sufficient redundancy make much difference. Each geographical site is unique and contains a combination of specific characteristics that distinguishes it from the other sites in a given series. Indeed, as the parapsychologists themselves have asserted, unless this were so, there would be no possibility of the transcripts' being uniquely associated with a given target site. In every one of the remote viewing experiments that allows the possibility of subtle cueing, the possibility of the judges' being able to make completely successful matchings because of this artifact is highly plausible; and as long as a highly plausible, normal alternative to ESP can account for the apparent success of the outcomes the parapsychologists, by their own standards, cannot claim evidence for paranormal transmission of information.

As it turns out, all but one of the nine scientifically reported studies of remote viewing (at the time of the Targ and Harary survey) suffer from the flaw of sensory cueing. The one experiment that cannot be faulted for this reason is the long-distance remote viewing experiment of Schlitz and Gruber (1980). However, as Hyman (1984–1985) has pointed out, this experiment suffers from another very serious flaw. Gruber, who was a member of the target team and thus was familiar with the targets, translated the subject's target descriptions into Italian for the judging process. Why the experimenters allowed such potential sources of biased experimental procedures is not known, but the violation obviously negates the results as evidence for psi.

Since the Targ and Harary survey, we have learned of two attempts

to replicate the Schlitz and Gruber experiment without the flaw mentioned. One, still unpublished, produced negative results. The second, by Schlitz and Haight (1984), produced marginally significant results. Indeed, if the more acceptable two-tailed test of significance had been used, the results would not have been considered significant by customary standards. Although the report of this study lacks sufficient documentation with respect to certain aspects of procedure, both Palmer (1985) and Alcock agree that this is the best controlled and most methodologically sound of all the remote viewing experiments so far.

In summary, after approximately 15 years of claims and sometimes bitter controversy, the literature on remote viewing has managed to produce only one possibly successful experiment that is not seriously flawed in its methodology—and that one experiment provides only marginal evidence for the existence of ESP. By both scientific and parapsychological standards, then, the case for remote viewing is not just very weak, but virtually nonexistent. It seems that the preeminent position that remote viewing occupies in the minds of many proponents results from the highly exaggerated claims made for the early experiments, as well as the subjectively compelling, but illusory, correspondences that experimenters and participants find between components of the descriptions and the target sites.

Research on Random Number Generators

The Basic Paradigm

The use of random number (or random event) generators for parapsychological research began in the 1960s and became relatively standard during the 1970s as the technology became widely available. A random number generator (RNG) is simply an electronic device that uses either radioactive decay or electronic noise to generate a sequence of random symbols. Originally such devices were used to test ESP, usually clairvoyance or precognition, but the most widespread and widely known work focuses on what is called micropsychokinesis, or micro-PK. In such research a subject, or operator, attempts to mentally bias the output of the random number generator, so that it produces a nonrandom sequence.

Most of the work with RNGs has used binary generators, or what Schmidt calls "electronic coin flippers." The output on each trial is either 0 or 1, that is, heads or tails. If the RNG is unbiased and truly random, then it should produce, on control runs, sequences of 0s and 1s that are independent of each other and that, in the long run, will yield 1s 50 percent of the time.

In a typical experiment, a subject (either a person who claims to be a psychic or a person chosen for availability who does not make such claims) is placed in the vicinity of the RNG and attempts to bias the output either toward more or fewer 1s. When an animal is used as the subject, the RNG output is usually coupled to an outcome whose frequency the animal presumably would like to either increase or decrease. In an experiment carried out with cockroaches, for example, one outcome was electric shock. If, during the time the output of the RNG was coupled with the shock apparatus, the proportion of shocks decreased below 50 percent, this would be taken as evidence of a psychokinetic effect of the cockroach on the output of the RNG.

The RNG experiments have been of interest to some military and governmental personnel because of the possibility, if such micro-PK is demonstrable, of psychically affecting equipment and computers that depend on the output of electronic symbols.

Results of the Experiments

In a recent survey 56 reports published between 1969 and 1984 and dealing with research on possible psychokinetic perturbations of binary RNGs (Radin, May, and Thomson, 1985), the reviewers counted 332 separate experiments. Of the 332 experiments, 188 were reported in refereed journals or conference proceedings, and of these 188 experiments with some claim to scientific status, 58 reported statistically significant results (compared with the 9 or 10 experiments that would be expected by chance). The other 144 experiments were produced by the Engineering Anomalies Research Laboratory at Princeton University; none of them had been published in a refereed journal at the time of the survey. Of these 144 experiments, 13 were classified as yielding statistically significant results. So, in the total sample of 332 experiments, 71 yielded ostensibly significant results at the traditional .05 level. This amounts to a success rate of approximately 21 percent, compared with the rate of 5 percent that would be expected by chance.

Palmer (1985) and Alcock agree that such results cannot be accounted for by chance. In other words, both the parapsychologist and the skeptic, in their respective reviews of the RNG research, agree that something other than accidental fluctuation is producing these results. Palmer calls this something an anomaly, which, while it may or may not be paranormal, cannot be explained by current scientific theories. Alcock points to various defects in the experimental protocols and concludes that no conclusions about the origins of these departures from randomness are justified until successful

outcomes can be more or less consistently produced with adequately designed and executed experiments.

Both Palmer and Alcock focus their reviews on the two most influential research programs on RNGs. One is the program of Helmut Schmidt, a quantum physicist who began working on psi and RNGs in 1969. The other is the program begun by Robert Jahn in the late 1970s, when he was dean of the School of Engineering and Applied Science at Princeton University (see Jahn, 1982). These two programs have accounted for almost 60 percent of all known experiments on RNGs. They have also been the most consistently successful in achieving statistically significant outcomes.

Although the results suggest that on each experimental group of trials the number of 1s is greater or less than the 50 percent baseline (depending on the intended direction), the actual degree of deviation from chance is quite small. As Palmer (1985) indicates, Schmidt's subjects have averaged approximately 50.5 percent hits over the years, compared with the expected baseline of 50 percent. This amounts to producing one extra 1 every 100 trials. The reason such a small departure from chance is statistically significant is that an enormous number of trials is conducted with each subject.

Jahn and his colleagues at Princeton have, in a much shorter time, produced on the order of 200 times the number of trials that Schmidt did in 17 years. The Princeton researchers have also produced a significantly lower success rate than Schmidt. In their formal series of 78 million trials, the percentage of hits in the intended direction was only 50.02 percent, or an average of 2 extra hits every 2,500 trials. Again, such an extremely weak effect is statistically significant only when one is dealing with very large numbers of trials.

Scientific Assessment of the RNG Experiments

Palmer (1985) carefully reviews the major criticisms of the work of Schmidt and Jahn. He addresses questions about security, because subjects often are left alone with the apparatus during the data collection. In the Princeton experiments, the data are always collected when the subject is alone with the apparatus. Although the Princeton experiments now contain a number of features that would make it extremely difficult for a naive subject to bias the results, it is not clear that this has always been so. It would make good scientific sense to conduct some trials during which the subject is carefully monitored to see if successful outcomes are still obtained.

The major reservations about the RNG experiments concern the adequacy of the randomization of the outputs. Schmidt applied only limited tests for the randomness of his machines, and most of the

control trials were gathered by allowing the machine to run for long periods, usually overnight. Although these controls usually produced results in line with the chance baseline, critics have pointed out that the controls are unsatisfactory because they were not conducted for shorter runs and at the same time as the data from the experimental sessions.

Palmer grants that the critics are correct in pointing out some of the shortcomings in Schmidt's methods for testing and controlling for the randomization of his machines. Palmer also correctly points out that such criticism is somewhat blunted by the fact that the critics have not specified any plausible mechanisms that would account for the obtained differences between the experimental and control trials. He is correct in pointing out that the Princeton experiments provide more adequate controls; however, he has probably assumed that the baseline controls in the Princeton experiments were run at the same time as the two experimental conditions of hitting and missing. It is easy to interpret the somewhat ambiguous description of the procedure in this manner. The relevant part of the authors' methodological description is as follows (Nelson, Dunne, and Jahn, 1984:9):

> The primary variable in these experiments is the operator's pre-recorded intention to shift the trial counts to higher or lower numbers. This directional intention may be the operator's choice—the so-called "volitional" mode—or it may be assigned by a specified random process—the "instructed" mode. In either mode, data are collected in a "tri-polar" protocol, wherein trials taken under an intention to achieve high numbers (PK+), trials taken under an intention to achieve low numbers (PK−), and trials taken as baseline, i.e. under null intention (BL), are interspersed in some reasonable fashion, with all other operating conditions held identical. For all three streams of data, effect size is measured relative to the theoretical chance mean. This tri-polar protocol is the ultimate safeguard in precluding any artifacts such as residual electronic biases or transient environmental influences from systematically distorting the data.

At first glance it might appear as if the tripolar protocol requires that the two types of experimental groups of trials and the baseline group of trials always be taken at the same session. This would be consistent with the claim that "any artifacts such as residual electronic biases or transient environmental influences" were thereby precluded "from systematically distorting the data." Such a claim would be justified if, in fact, at each session one group of trials of each of the three types was obtained, provided that each group of trials was of the same length and that the order of the three types of trials was independently randomized for each session.

The description provided by Nelson and his colleagues says nothing

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

at all about the order in which the three conditions were conducted, and a careful reading indicates that the baseline data may not always have been obtained at the same sessions and under the same conditions as the experimental groups of trials. It is not clear what the authors mean by stating that the three trials "are interspersed in some reasonable fashion." In fact, an examination of the data reported for each subject makes it clear that the strict tripolar protocol could not possibly have been followed with much of the data collection, because in many cases the baseline data are entirely absent or occur with many fewer trials than the experimental data. Indeed, it is not even clear that PK+ and PK− trials were always obtained at the same sessions, because for some subjects the total numbers of these trials are not equal.

We suspect that, over the six years or so during which the Princeton group was accumulating its data base, it made many changes in both the hardware and the experimental protocol. The sophisticated procedures currently in use and the requirement that the three types of trials be of equal length and that one of each be conducted at each session are the most recent variations in the paradigm. Unfortunately, the data are not presented in such a way that it is possible to determine whether the successful results are due to the earlier or the later experiments.

Such issues become especially important when we consider the extremely small size of the effect being claimed and when we further realize, as Palmer has pointed out, that the bulk of the significance in the formal series was due to just one subject, who contributed 23 percent of the total data. This one subject achieved a hit rate of 50.05 percent. When her data are eliminated, the remaining data yield a hit rate of 50.01 percent, which is no longer significantly different from chance.

In other words, it looks as if almost all the success of Jahn's huge data base can be attributed to the results from one individual, who, over the years, produced almost 25 percent of the data. This one individual was not only the most experienced subject, but also, presumably, familiar with the equipment. When combined with the fact, as Palmer points out, that the Princeton experiments provide inadequate documentation on precautions to prevent tampering by subjects, it becomes even more important to see if the same degree of success can be achieved when the sessions are adequately monitored.

Alcock, in his review of the same RNG studies surveyed by Palmer, points to a number of weaknesses in both the Schmidt and the Princeton experiments. For example, he faults Schmidt's experiments for such things as inadequate controls, failure to examine the target sequences, overcomplicated experimental setups, inadequate tests of randomness, and lack of methodological rigor. Alcock faults the Princeton experiments for such things as failing to randomize the sequence of groups of trials at each session, inadequate documentation on precautions against data tampering, and possibilities of data selection.

Palmer and Alcock do not really differ in their assessments of the shortcomings of the Schmidt and Princeton RNG experiments. They do differ, however, on what conclusions can be drawn from such imperfect experiments. Palmer emphasizes the fact that the critics have not provided plausible explanations as to how the admitted flaws could have caused the observed results. His position seems to be that, unless the critics can provide such plausible alternatives, the results should be accepted as demonstrating an anomaly. Alcock focuses on the fact that the successful results have been obtained under conditions that fall short of the experimental ideals that parapsychologists themselves profess. He emphasizes that the parapsychologists have no right to claim to have demonstrated psi from experiments that have been conducted with "dirty test tubes." Such a revolutionary conclusion as the existence of psi demands justification from experiments that have clearly used "clean test tubes."

What would it take to conduct an adequate RNG experiment? May, Humphrey, and Hubbard (1980) set out to do just that. After reviewing all available RNG experiments from 1970 through 1979 and taking into account the various deficiencies in these experiments, they gathered together and meticulously tested the components necessary to provide adequately randomized trials. They also devised a careful experimental protocol and set out in advance the precise criteria that would have to be fulfilled before they could call their results successful. Going further, after they completed the experiment with results that met their criteria for success, they subjected their equipment to all sorts of physical extremes to see if they could obtain such a degree of success by a possible artifact.

They report that this singularly well controlled RNG experiment in fact met their criteria for success. It is unfortunate, therefore, that this carefully thought-out experiment was conducted only once. After the one successful series, using seven subjects, the equipment was dismantled, and the authors have no intention of trying to replicate it (personal communication, August 1986). It is unfortunate because this appears to be the only near-flawless RNG experiment known to us, and the results were just barely significant. Only two of the seven subjects produced significant results, and the test of overall significance for the total formal series yielded a probability of 0.029.

The experiment, while nearly flawless, still had some problems as evidence for psi. For one thing, it was reported only in a technical report in 1980 and has never been published in a refereed scientific journal. Despite the admirable attention to details, all the control trials were taken when no human being was present. One might argue that this was not an ideal control for the experimental session, in which a subject was physically present in the room. The authors have assured us that their various attempts to bias the machine by physical means almost certainly rule out the possibility that the mere presence of a human being could have affected the output. However, a physicist who claims to have several years of experience in constructing and testing random number devices tells us that it is quite possible, under some circumstances, for the human body to act as an antenna and, as a result, possibly bias the output.

May and his colleagues at SRI, in the same technical report in which they claim successful results for their single experiment, surveyed all the RNG experiments known to them through the year 1979 and found that their combined significance was astronomically high. They add (May, Humphrey, and Hubbard, 1980:8):

> This impressive statistic must, however, be evaluated with respect to experimental equipment and protocols. All the studies surveyed could be considered incomplete in at least one of the following four areas: (1) No control tests were reported in more than 44 percent of the references. Of those that did, most did not check for temporal stability of the random sources during the course of the experiment. (2) There were insufficient details about the physics and constructed parameters of the experimental apparatus to assess the possibility of environmental influences. (3) The raw data was not saved for later and independent analysis in virtually any of the experiments. (4) None of the experiments reported controlled and limited access to the experimental apparatus.

As far as we can tell, the same four points can be made with respect to the RNG experiments that have been conducted since 1980. The situation for the RNG experiments thus seems to be the same as that for remote viewing: over a period of approximately 15 years of research, only one successful experiment can be found that appears to meet most of the minimal criteria of scientific acceptability, and that one successful experiment yielded results that are just marginally significant.

Research on the Ganzfeld

The Ganzfeld Experiments

The Ganzfeld psi experiments are named after the term used by Gestalt psychologists to designate the entire visual field. For theoretical purposes, the Gestalt psychologists wanted to create a situation in which the subject or observer could view a homogeneous visual field, one with no imperfections or boundaries. Psychologists later discovered that when individuals are put into a Ganzfeld situation they tend quickly to experience what they described as an altered state of mind.

In the early 1970s, some parapsychologists decided that the use of the Ganzfeld would provide a relatively safe and easy way to create an altered state in their experimental subjects. They believed that such a state was more conducive to picking up the elusive psi signals. In a typical psi Ganzfeld experiment, the subject, or percipient, has halved ping-pong balls taped over the eyes. The subject then reclines in a comfortable chair while white noise plays through earphones attached to his or her head. A bright light shines in front of the subject's face. When seen through the translucent ping-pong balls, the light is experienced as a homogeneous, foglike field. When so prepared, almost all subjects report experiencing a pleasant, altered state within 15 minutes.

While one experimenter is preparing the subject for the Ganzfeld state, a second experimenter randomly selects a target pool from a large set. The target pool typically consists of four possible targets, usually reproductions of paintings or pictures of travel scenes. One of the four is chosen at random to be the target for that trial. The target is given to an agent, or sender, who tries to communicate its substance psychically to the subject in the Ganzfeld state. After a designated period, the subject is removed from the Ganzfeld state and presented with the four candidates from the target pool. The subject then ranks the four candidates in terms of how well each matched the experience of the Ganzfeld period. If the actual target is ranked first, the trial is designated a hit. An actual experiment consists of several trials. In the example, the probability is that one of every four trials will produce a hit. If the number of hits significantly exceeds the expected 25 percent, then the result is considered to be evidence for the existence of psi.

Critique of the Ganzfeld Experiments

In a careful and systematic review of the Ganzfeld experiments undertaken in 1981 and published in the March 1985 issue of the *Journal of Parapsychology*, Hyman concluded that the data base exhibited flaws involving multiple testing, inadequate controls for sensory leakage, inadequate randomization, statistical errors, and inadequate documentation. These flaws, in his opinion, were sufficient

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

to disqualify the Ganzfeld data base as evidence for psi. Of the 42 experiments, 39 (93 percent) used multiple analyses, which artificially inflated the chances of obtaining significant outcomes. Only 11 (26 percent) clearly indicated that they had adequately randomized the target selections. As many as 15 (36 percent) used inferior randomization, such as hand shuffling, or no randomization at all. The remaining 16 experiments did not supply sufficient information on how they had chosen the targets. As many as 23 of the experiments (55 percent) used only one target pool, which means that the subject was handed for judging not a copy of the target but the very same target that the percipient had handled, permitting the possibility of sensory cueing. Although the argument for psi is mainly a statistical one, the reports of 12 experiments (29 percent) revealed statistical errors. A number of other departures from optimal practice were also found.

The same issue of the *Journal of Parapsychology* contained a lengthy rebuttal by parapsychologist Charles Honorton, one of the pioneers of the Ganzfeld psi technique. Honorton disputed many of Hyman's opinions as to what constituted flaws; provided a reanalysis of the data base to overcome many of the statistical weaknesses of the original experiments; and argued that the flaws he agreed existed were not sufficient to have accounted for the findings. In this respect his analysis is consistent with Palmer's approach. He does not deny that the experiments depart from optimal design, but he argues that such departures are insufficient to account for the results.

Honorton and Hyman had the opportunity to discuss their differences about psi in general at the Parapsychological Association meetings in 1986; as a result, they agreed to draft a joint communiqué to emphasize those points on which they agree. That communiqué appeared in the December issue of the *Journal of Parapsychology* (Hyman and Honorton, 1986). They agree that the current data base is insufficient to support either the conclusion that psi exists or the conclusion that the results are due to artifacts. They further agree that the issue can be settled only by future experiments conducted according to the stated standards of parapsychology, which are also the accepted standards of psychological research.

Another important input to the committee's judgment on the Ganzfeld research was the systematic evaluation of the contemporary parapsychological literature by Charles Akers (1984), a former parapsychologist. Akers's critique used a methodological strategy different from that used by Hyman. Hyman undertook to evaluate the entire data base of a single research paradigm (Ganzfeld), including both successful and unsuccessful outcomes. Akers surveyed

contemporary ESP experiments broadly, but confined his evaluation to those that had produced significant results with unselected subjects. Hyman assigned flaws to experiments without regard to whether each flaw, by itself, could have caused the observed outcome. Akers charged a flaw to a study only if he thought the flaw could have been sufficient to produce the observed result. He chose a sample of 54 parapsychological experiments from areas of research that had been previously reviewed by Honorton or Palmer; his intent was to choose experiments that could be viewed as the best current evidence for the existence of psi. As a result of this exercise, he concluded (Akers, 1984:160–161):

> Results from the 54-experiment survey have demonstrated that there are many alternative explanations for ESP phenomena; the choice is not simply between psi and experimenter fraud. . . . The numbers of experiments . . . flawed on various grounds were as follows: randomization failures (13), sensory leakage (22), subject cheating (12), recording errors (10), classification or scoring errors (9), statistical errors (12), reporting failures (10). . . . All told, 85% of the experiments were considered flawed (46/54).
>
> This leaves eight experiments where no flaws were assigned. . . . Although none of these experiments has a glaring weakness, this does not mean that they are especially strong in either their methods or their results. . . .
>
> In conclusion, eight experiments were conducted with reasonable care, but none of these could be considered as methodologically ideal. When all 54 experiments are considered, it can be stated that the research methods are too weak to establish the existence of a paranormal phenomenon.

Research on Electrical Activity and Emotional States

The Backster Laboratory

In addition to examining parapsychological research in areas that have produced large literatures, the committee witnessed an example of experimental work at a far less developed stage. On February 10, 1986, committee members visited the Backster Research Foundation in San Diego and saw a demonstration of experimental procedures for detecting a correlation between the electrical activity of oral leukocytes and the emotional states of the donor.

Cleve Backster is a polygraph specialist who had at one time helped develop interrogation techniques for the Central Intelligence Agency and now runs his own polygraph school in San Diego. The school is housed in the same rooms that constitute the Backster Research Foundation, which is devoted to the study of what Backster refers to as primary perception. Backster's research on paranormal matters

began in February 1966, when he recorded, from a philodendron plant that he had hooked up to a polygraph, a response he recognized as similar to that of human beings in emotional states. Backster believed he had demonstrated that the plant showed such emotional response when brine shrimp or other living organisms were either threatened or actually killed in an adjoining room. The notion of primary perception in plants became both a popular subject for research and a highly controversial concept during the late 1960s and early 1970s.

We were told that Backster has quietly continued his researches into this and related matters. He has now devised a technique for recording electrical activity in leukocytes taken from a donor's mouth. The advantage of this technique, we were told, is that the leukocytes respond mostly to emotional states of the donor.

One committee member volunteered to be the demonstration subject. Another member accompanied him to observe the techniques for obtaining the leukocytes and preparing them for recording. The sample was obtained by having the subject "chew" on a 1.2 percent saline solution and then spit it back into a centrifuge tube. Ten such samples were obtained in this way. The samples were then spun in a centrifuge for six minutes, and the particulate matter at the bottom of each tube was pipetted into the preparation tube. The preparation tube contained about one centimeter of particulate matter and was filled almost to the top with 1.2 percent saline solution. Two uninsulated wire electrodes were inserted into the bottom of the tube, which was then placed within a shielded cage and connected by leads to an EEG-type recording apparatus.

During the demonstration, the subject sat approximately two meters from the preparation. We were told that subjects usually sit about five meters from the preparation. A split-screen projection video display was provided: the lower portion of the screen recorded the movements of the polygraph paper and pen as they produced a record of the electrical activity presumably taking place in the leukocyte preparation. The upper portion of the screen recorded the behavior of the seated subject.

In his previous research using this arrangement, Backster reported that, when the subject revealed an emotional reaction, the electrical action of the leukocytes showed a corresponding reaction. During our demonstration, the polygraph record produced several strong deflections in both the control and the experimental series, but they did not obviously correlate with any corresponding thoughts or emotional states of the subject as various stimuli were presented. Backster suggested that this was probably because so many people were crowded into the laboratory that the leukocytes were responding to thoughts and feelings of other individuals in the room. Thus, a demonstration of results, as opposed to techniques, was not, after all, going to be possible during our visit.

Backster then showed us videotapes of the split-screen results he had obtained in his "formal" experiments. The results consisted of 12 examples of apparent correlations between an emotional response and a deflection of the polygraph record. The 12 examples came from 7 sessions with 7 different subjects. Although the information is not given in his written report, it appears that each session lasted for approximately half an hour. During this time, the donor is engaged in conversation or watches videotapes of television programs. The sessions are not standardized or planned. Backster's intent, apparently, is to elicit spontaneous emotional responses from a subject during the session. He believes that a stimulus that evokes an emotional response in one subject will not necessarily do so in another subject.

In one example, the subject was a young man who was looking at an issue of *Playboy* magazine. The polygraph tracing began to display large deflections soon after he encountered a nude photograph of an attractive young woman. The large deflections continued for approximately two minutes; the tracing slowly settled down to normal activity after the magazine was closed. Soon after, the young man reached for the closed magazine, and the record reveals a single deflection at that point. In another example, the subject was a retired police lieutenant. When discussing his approaching retirement, he was asked a question about his wife's attitude toward having him "underfoot." A large deflection of the polygraph tracing occurred soon after this question was asked. When asked, the donor confirmed that he was emotionally aroused at that moment in the session (see Backster and White, 1985).

Cleve Backster and his supporters apparently believe that he has successfully demonstrated that detached oral leukocytes respond to the emotions of their donor even when separated by as much as several miles. They also believe that these results are reliable and replicable.

Critique of the Backster Experiment

What we have read and observed about Backster's procedures does not justify the claim he is making. His answers to our questions made it clear that he has not considered using the appropriate controls needed to ensure that the obtained "correlations" are real and due to the causes he has assumed. To make adequate physiological recordings from a

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

preparation of in vitro leukocytes and to demonstrate the correlation between emotional response and leukocyte activity requires experimental arrangements and procedures at a level of sophistication well beyond those we observed.

Committee members who are knowledgeable about the procedures and instrumentation of psychophysiological experiments expressed doubts about the adequacy of the setup to perform the tasks Backster has undertaken. Serious doubts were expressed about the possibility that the leukocytes were alive at the time of recording. Further doubts were expressed about the setup's ability to avoid contamination of the recording procedures by stray influences of various sorts. We do not discuss these drawbacks in detail here. We confine our discussion to Backster's method for establishing a correlation between the alleged activity of the detached leukocytes and the emotional state of the donor. When we consider how the existence of such correlations was established, we again see how inappropriate methodology can lead to very misleading conclusions.

Many problems exist with regard to Backster's procedures for detecting correlations. In trying to demonstrate a pattern of covariation between two records of behavior over time, one record is the tracing of amplified electrical activity coming from the electrodes and through the leads. Although this tracing can be quantified, Backster has apparently made no attempt to do so. Instead, he has relied on visual inspection of the polygraph record to pick out points at which the deflections of the pen from the baseline are noticeable. Although such subjective judgment is scientifically unacceptable, the deflections that he uses in his examples seem sufficiently marked that they probably can be considered to be real deviations from the baseline. At any rate, let us assume that responses on the polygraph record can be visually pinpointed with reasonable objectivity.

The deflections on the polygraph record are then compared with happenings on the concurrent videotaping of the conversation with the subject. Here we encounter very serious problems as to what constitutes an emotional response on this behavioral record. Backster believes he can identify categories of potentially emotionally arousing stimuli in the nonstandardized, qualitative, ongoing record of conversation. He then can determine if the subject was experiencing an emotional reaction to such a stimulus by simply replaying the record, pointing to the segment that corresponds to a place where the polygraph showed a deflection, and asking the subject if he or she recalls what was taking place at that moment as an emotionally arousing experience. If the subject agrees, this is said to confirm a "correlation" between the emotional state and the corresponding activity of the tracing.

Such a purely subjective determination of an emotional response opens

the process to a variety of known biases, many of them discussed in the paper prepared for the committee by Griffin (Appendix B). The literature on "illusory correlation" (Alloy and Tabachnik, 1984; Griffin paper) makes it clear how subjective expectations and cognitive biases can lead to false impressions of correlation. Backster's method of searching for correlations compounds these inevitable biases: he does not independently determine moments of emotional response in the subject's behavioral record and moments of polygraph deflections and then look for a match between the two. Instead, he apparently looks for polygraph deflections and then tries to determine if an emotional response can be found that occurred in the vicinity of the polygraph activity. In other words, the determination of the emotional response is done with full knowledge of the fact that a polygraph deflection has occurred.

Under such circumstances, we would expect processes of subjective validation to operate. In addition, the method of verifying the emotional response, by asking the subject to acknowledge that he or she was in fact experiencing such a state at the moment the polygraph record indicated a leukocyte response, is itself suspect. This is the sort of circumstance in which demand characteristics (i.e., responses determined by the presumed intent of the experimenters) are known to operate.

Good science dictates that the moments of emotional response should be determined independently of the moments of polygraph response. Both the experimenter and the subject must be blind to the polygraph record when determining the moments of emotional response. Only when the determination of events on the two records has been made independently of each other can the records be compared to determine if the emotional responses and the polygraph activity are correlated.

Illusory correlations occur because our subjective judgments of covariation tend to use only a portion of the relevant information and because we tend to bias observed events in terms of our expectations. In particular, intuitive judgments of covariation tend to focus only on the co-occurrence of treatment of interest and successful outcomes, ignoring times when the treatment co-occurred with unsuccessful outcomes. Backster uses only those examples from his records in which an emotional response co-occurs with a polygraph deflection; the 12 such examples from the 7 experimental series represent a very small fraction of the total data collected.

Not only is a sample of just 12 co-occurrences probably too small for estimating whether a true correlation exists, but it is also impossible from this information alone to estimate whether any correlation exists. All the data are needed for this purpose. Almost certainly, more than 12 polygraph deflections must have appeared in the total record. In the brief demonstration for the committee, both the control and the experimental series

yielded several deflections, so it is reasonable to assume that many more than 12 deflections were obtained in the complete record. It is likely that these unreported deflections were not preceded by any emotional responses.

Almost certainly, more than 12 emotional responses must have appeared in the total record. The point of conducting the sessions was to expose the subjects to a variety of emotional stimuli; therefore, it is essential to know the number of times that emotional responses occurred *without* the corresponding occurrence of polygraph responses. Finally, to determine correlation, it is essential to know the frequency of co-occurrence of the absence of emotional responses and the absence of polygraph responses. All this information is needed to determine whether the claimed correlation exists. All the data must be used. From these data, one can compare the proportion of times that an emotional response is followed by a polygraph response with the proportion of times that the absence of an emotional response is followed by a polygraph response. Only if these two proportions are significantly different from one another can we assume that the data provide evidence for a correlation between emotional response and leukocyte activity. The fact that Backster was able to find 12 examples of the co-occurrence between emotional response and polygraph deflection, even if these correspondences had come from double-blind matching, provides us with absolutely no information about whether a correlation exists.

The stronger claim would be, of course, not that a correlation exists, but that a causal connection exists between the subject's emotional states and the responses of the detached leukocytes. As Chapter 3 on evaluation indicates, such a causal explanation requires much more than the demonstration of correlation between two series. Because Backster did not use double-blind procedures to determine emotional responses, and because the procedures he did use are known to be just those that facilitate the occurrence of a variety of subjective biases, he may well have obtained a correlation between his two series. However, his procedures for finding such correlations are sufficiently flawed that we do not know if in fact the suspected (and presumably biased) correlation actually does exist in his data. The Backster experiment indicates that the best intentions combined with scientific instrumentation and polygraphic records cannot, in themselves, guarantee data of scientific quality.

Discussion of the Scientific Evidence

Both the parapsychologists cited in this report and the critics of parapsychology believe that the best contemporary experiments in parapsychology fall short of acceptable methodological standards. The critics

conclude that such data, based on methodologically flawed procedures, cannot justify any conclusions about psi. The parapsychologists argue that, while each experiment is individually flawed, when taken together they justify the conclusion that psi exists.

Palmer's conclusion in this regard is unique. Although he agrees that the data do not justify the conclusion that a paranormal phenomenon has been demonstrated, he argues that the data, with all their drawbacks, do justify the conclusion that an anomaly of some sort has been demonstrated. It is this purported demonstration of an anomaly that, according to Palmer, further justifies the claim that parapsychologists do have a subject matter. The awkward aspect of Palmer's position is that, without an adequate theory, there is no way to know that the anomaly "demonstrated" in one experiment is the same anomaly "demonstrated" in another; indeed, there is no limit to the possible causes of the anomaly in a given experiment. Without an adequate theory, there is no reason to assume that the various anomalies constitute a coherent or intelligibly related class of phenomena.

The committee distinguishes among three types of criticism that can be leveled at a given parapsychological finding. The first is what we might refer to as the smoking gun. This type of criticism asserts or strongly implies that the observed findings were due not to psi but to factor X. Such a claim puts the burden of proof on the critic. To back up such a claim, the critic must provide evidence that the results were in fact caused by X. Many of the bitterly contested feuds between critics and proponents have often been the result of the proponent's assuming, correctly or incorrectly, that this type of criticism was being made.

The second type of criticism can be referred to as the plausible alternative. In this case, the critic does not assert that the result *was* due to factor X, but instead asserts that the result *could have been* due to factor X. Such a stance also places a burden on the critic, but one not so stringent as the smoking gun assertion. The critic now has to make a plausible case for the possibility that factor X was sufficient to have caused the result. For example, optional stopping of an experiment on the part of a subject can bias the results, but the bias is a small one; it would be a mistake to assert that an outcome was due to optional stopping if the probability of the outcome is extremely low. Akers's critique, which was previously discussed, is an example based on the plausible alternative.

The third type of criticism is what we have called the dirty test tube. In this case, the critic does not claim that the results have been produced by some artifact, but instead points out that the results have been obtained under conditions that fail to meet generally accepted standards. The gist of this type of criticism is that test tubes should be clean when doing

careful and important scientific research. To the extent that the test tubes were dirty, it is suggested that the experiment was not carried out according to acceptable standards. Consequently, the results remain suspect even though the critic cannot demonstrate that the dirt in the test tubes was sufficient to have produced the outcome. Hyman's critique of the Ganzfeld psi research and Alcock's paper on remote viewing and random number generator research are examples of this type of criticism.

In the committee's view, it is in this latter sense, the dirty test tube sense, that the best parapsychological experiments fall short. We do not have a smoking gun, nor have we demonstrated a plausible alternative; but we imagine that even the parapsychological community must be concerned that their best experiments still fall far short of the methodological adequacy that they themselves profess.

Honorton and Hyman differ on whether to assign a flaw in randomization to a particular series of experiments. With Honorton's assignment, the studies with adequate randomization do not differ in significance of outcome from those with inadequate randomization. With Hyman's assignment, the experiments with inadequate randomization have significantly more successful outcomes than do those with adequate randomization. A simple disagreement on one experiment can thus make a huge difference as to whether we conclude that this flaw contributed or did not contribute to the observed outcomes. Several similar examples could be cited to illustrate the extreme sensitivity of this data base to slight changes in flaw assignments.

Even if Palmer is correct in asserting that in a particular case an anomaly has been demonstrated, serious problems remain. In astronomy and other sciences, an anomaly is a very precise and specifiable departure from a well-defined theoretical expectation. Neptune was discovered, for example, when Leverrier was able to specify not only that the orbit of Uranus departed from that expected by Newtonian theory, but also precisely in what way it departed from expectation. Nothing approaching such a specifiable anomaly has been claimed for parapsychology. A vague and unspecifiable departure from chance is a far cry from a well-described and systematic departure from a precise, theoretical equation. Leverrier's anomaly was consistent with only a very narrow range of possibilities. The sort of anomaly claimed for parapsychology is currently consistent with an almost infinite variety of possibilities, including artifacts of various kinds.

THE PROBLEM OF QUALITATIVE EVIDENCE

The committee continually encountered the distinction between qualitative and quantitative evidence for the existence of paranormal phenomena. Many proponents of the paranormal acknowledge such a difference in one way or another. Some realize that it is only quantitative evidence that will convince the scientific community. Although they themselves have relied on qualitative evidence for their own beliefs, they refer us to the RNG experiments of Robert Jahn or the remote viewing experiments at SRI as examples of supporting quantitative data.

Most proponents seem impatient with the request for scientific evidence. They have been convinced through their own experiences or the vivid testimonies of individuals whom they trust. Many argue that qualitative evidence can be as good as quantitative; indeed, they claim that in some circumstances it can be better.

The arguments for the superiority of qualitative evidence are based in many cases on such factors as ecological validity, conducive atmosphere, and holism. The ecological validity argument asserts that the artificial conditions required for laboratory experiments are so different from the natural settings in which paranormal phenomena typically occur that findings from such controlled studies are irrelevant. By removing the psychic from his or her natural domain or by arranging conditions to suit the needs of scientific observation, it is claimed, the scientist destroys the very phenomenon under question. The ecological validity argument is closely related to the other arguments. Proponents who emphasize the conducive atmosphere assert that the austere conditions of strict laboratory procedure create an atmosphere that is numbing or inimical to psychic functioning. Those who emphasize holism point out that the experimental procedures necessarily dissect and focus on restricted portions of a system. Such compartmentalization, it is claimed, makes it impossible to study the sorts of paranormal phenomena that operate only as a total system in a naturalistic context.

Qualitative Evidence and Subjective Biases

What is meant by qualitative evidence? Roughly, it means any sort of nonscientific evidence that proponents find personally convincing. Typically, it involves personally experiencing or witnessing the phenomenon. Less compelling, but still effective, is the testimony of friends or trusted acquaintances who have personally experienced it. Even individuals who are intellectually aware of the pitfalls of personal observation and testimony find it difficult, even impossible, to disregard the compelling quality of such evidence in the formation of their own beliefs.

A major parapsychologist admitted to one committee member that the scientific evidence did not justify concluding that psi exists. "As a trained scientist," he said, "I know quite well that by scientific criteria there is no evidence for the existence of psi. In fact, I have always argued with

Approved For Release 2002/05/17 : CIA-RDP96-00791R000200320001-1

my parapsychological colleagues that they are making a serious mistake in trying to get the scientific community to take their current evidence seriously. Before they do this, they first have to be able to collect the sort of repeatable and lawful data that constitute scientific evidence." This same parapsychologist then explained why, despite the current lack of evidence, he remained a parapsychologist. "When I was 16 I had some personal experiences of a psychic nature that were so compelling that I have no doubt that they were real. Yet, as a trained scientist, I know that my personal experiences and subjective convictions cannot and should not be the basis for asking others to believe me." This parapsychologist is unusual in that he makes the distinction within himself between beliefs that are subjectively compelling and beliefs that are scientifically justifiable. More typical is the proponent who, as a result of compelling personal experience, not only has no doubt about the reality of underlying paranormal cause, but also has no patience with the refusal of others to support that belief.

We see two problems regarding qualitative evidence. First, personal observation and testimony are subject to a variety of strong biases of which most of us are unaware. When such observations and testimony emerge from circumstances that are emotional and personal, the biases and distortions are greatly enhanced. Psychologists and others have found that the circumstances under which such evidence is obtained are just those that foster a variety of human biases and erroneous beliefs. Second, beliefs formed under such circumstances tend to carry a high degree of subjective certainty and often resist alteration by later, more reliable disconfirming data. Such beliefs become self-sealing, in that when new information comes along that would ordinarily contradict them, the believers find ways to turn the apparent contradictions into additional confirmation.

The committee asked Dale Griffin to describe many of the ways in which cognitive and social psychologists have documented that human subjective judgment can lead us astray. Griffin's paper emphasizes the cognitive biases termed *availability* and *representativeness*, but he also discusses motivational biases. Although most of these biases have been created under laboratory conditions, they are nonetheless quite powerful, and evidence has been mounting that, if anything, they are much more powerful in natural settings. Griffin points out that one vivid, concrete experience is usually sufficient to outweigh conclusions based on hundreds or thousands of cases based on abstract summary statistics. These and the other biases discussed by Griffin should make us wary of conclusions based on qualitative evidence.

Examples of Problematic Beliefs

In this section we discuss some examples of beliefs about paranormal phenomena that have been formed under conditions known to generate cognitive illusions and strong delusional beliefs. We attempt to make clear why we are skeptical of any evidence offered in support of the paranormal that does not strictly fulfill scientific criteria. We believe it is important to realize the power of such conditions to create strong but false beliefs.

In 1974 a group of distinguished physicists at the University of London observed renowned psychic Uri Geller apparently bend metallic objects and cause part of a crystal, encapsulated in a container, to disappear.

Impressed with what they saw, in 1975 these scientists contributed an article to *Nature* outlining their ideas about how to conduct successful parapsychological research (reprinted in Hasted et al., 1976). In their discussion they note that successful results depend on the relation among the participants and that phenomena are more likely to occur when all participants are in a relaxed state, all sincerely want the psychic to succeed, and "the experimental arrangement is aesthetically or imaginatively appealing to the person with apparent psychokinetic powers."

Hasted and his colleagues describe further desiderata. The psychic should be treated as one of the experimental team, contributing to an attitude of mutual trust and confidence that facilitates successful appearance of the allegedly paranormal effects. The slightest hint of suspicion on the part of the observers can stifle the occurrence of any phenomena. Observers should avoid looking for any particular outcome that interferes with the required relaxed state of mind and impedes paranormal powers. To help avoid the inhibiting effects of concentrated attention, participants should talk and think about matters irrelevant to the experiment at hand.

Acknowledging that these desiderata make it difficult to preclude trickery, Hasted and his colleagues express confidence that they can both create psi-conducive conditions and eliminate the possibility of being tricked (Hasted et al., 1976:194):

> It should be possible to design experimental arrangements which are beyond any reasonable possibility of trickery, and which magicians will generally acknowledge to be so. In the first stages of our work we did in fact present Mr. Geller with several such arrangements, but these proved aesthetically unappealing to him.

Although we may sympathize with the British physicists' desire to create conditions conducive to the appearance of genuine psychic powers, if such powers exist, we cannot fail to note the quandary that their efforts produce. In their quest for psi-conducive conditions, they have created guidelines that play into the hands of anyone intent on deceiving them.

The very conditions that are specified as being conducive to the appearance of paranormal phenomena are almost always precisely those that are conducive to the successful performance of conjuring tricks. One of the first rules the aspiring conjuror learns is never to announce in advance the specific outcome that he or she is going to produce. In this way onlookers will not know where and on what they should focus their attention and consequently will be less apt to detect the method by which the trick was accomplished. The authors' advice to avoid focusing on a predetermined outcome greatly facilitates the conjuror's task.

The insistence that the arrangements meet with the psychic's approval is by far the most devastating of these conditions. Geller will perform only if the conditions are "aesthetically pleasing." This amounts to giving the alleged psychic complete veto power over any situation in which he or she feels that success is not ensured. This in turn means that the psychic being tested, not the experimenters, is controlling the experiment. Surely the British physicists ought to realize the irony of their admission that all their experimental arrangements designed to preclude trickery turned out to be aesthetically unacceptable to Uri Geller.

Another example of beliefs generated in circumstances that are known to create cognitive illustions is macro-PK, which is practiced at spoon-bending, or PK, parties. The 15 or more participants in a PK party, who usually pay a fee to attend and bring their own silverware, are guided through various rituals and encouraged to believe that, by cooperating with the leader, they can achieve a mental state in which their spoons and forks will apparently soften and bend through the agency of their minds.

Since 1981, although thousands of participants have apparently bent metal objects successfully, not one scientifically documented case of paranormal metal bending has been presented to the scientific community. Yet participants in the PK parties are convinced that they have both witnessed and personally produced paranormal metal bending. Over and over again we have been told by participants that they know that metal became paranormally deformed in their presence. This situation gives the distinct impression that proponents of macro-PK, having consistently failed to produce scientific evidence, have forsaken the scientific method and undertaken a campaign to convince themselves and others on the basis of clearly nonscientific data based on personal experience and testimony obtained under emotionally charged conditions.

Consider the conditions that leaders and participants agree facilitate spoon bending. Efforts are made to exclude critics because, it is asserted, skepticism and attempts to make objective observations can hinder or prevent the phenomena from appearing. As Houck, the originator of the PK party, describes it, the objective is to create in the participants a

peak emotional experience (Houck, 1984). To this end, various exercises involving relaxation, guided imagery, concentration, and chanting are performed. The participants are encouraged to shout at the silverware and to "disconnect" by deliberately avoiding looking at what their hands are doing. They are encouraged to shout Bend! throughout the party. "To help with the release of that initial concentration, people are encouraged to jump up or scream that theirs is bending, so that others can observe." Houck makes it clear that the objective is to create a state of emotional chaos. "Shouting at the silverware has also been added as a means of helping to enhance the emotional level in a group. This procedure adds to the intensity of the command to bend and helps create pandemonium throughout the party."

A PK party obviously is not the ideal situation for obtaining reliable observations. The conditions are just those which psychologists and others have described as creating states of heightened suggestibility and implanting compelling beliefs that may be unrelated to reality. It is beliefs acquired in this fashion that seem to motivate persons who urge us to take macro-PK seriously. Complete absence of any scientific evidence does not discourage the proponents; they have acquired their beliefs under circumstances that instill zeal and subjective certainty. Unfortunately, it is just these circumstances that foster false beliefs.

Discussion of Qualitative Evidence

Our analysis of the evidence put before us indicates that even the most solidly based arguments for the existence of paranormal phenomena fall short of the currently accepted parapsychological standards. Even if the best evidence had been collected according to acceptable scientific standards, most proponents would have in fact remained convinced by personal experiences and data that clearly fall far short of scientific acceptability. We have looked at two examples to make clear why and in what ways such failures to meet acceptable standards render the corresponding arguments useless as evidence for the paranormal, even though they have created compelling and strongly held beliefs in those who have been exposed to them.

The examples illustrate how different ways of attempting to acquire evidence for paranormal phenomena can depart from adequate standards. These inadequacies become especially critical when we note that the conditions under which the alleged paranormal phenomena are supposed to occur are just those known to foster biases and false beliefs. The PK parties, while creating powerful beliefs in paranormal metal bending, clearly violate almost every principle for obtaining trustworthy data. These parties offer no standardization, no objective records, and no