Research Note 86-63

AN EVALUATIVE REPORT ON THE CURRENT STATUS OF PARAPSYCHOLOGY

John Palmer
Parapsychology Laboratory, University of Utrecht

ARI Scientific Coordination Office, Europe
Michael Kaplan, Chief

BASIC RESEARCH OFFICE
Milton S. Katz, Director

U. S. Army

Research Institute for the Behavioral and Social Sciences

MAY 1986

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the
Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

WM. DARRYL HENDERSON
COL, IN
Commanding

Research accomplished under contract for the
Department of the Army

Parapsychology Laboratory, University of Utrecht
Utrecht, The Netherlands

Technical review by
Steve Kronheim

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARI Research Note 86-63	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) AN EVALUATIVE REPORT ON THE CURRENT STATUS OF PARAPSYCHOLOGY		5. TYPE OF REPORT & PERIOD COVERED Final Report August 1984 to December 1985 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) John Palmer		8. CONTRACT OR GRANT NUMBER(s) DAJA 45-84-M-0405
9. PERFORMING ORGANIZATION NAME AND ADDRESS Parapsychology Laboratory University of Utrecht, Sorbonnelaan 16 Utrecht, The Netherlands		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2Q161102B74F
11. CONTROLLING OFFICE NAME AND ADDRESS Army Research Institute Scientific Coordination Office, Europe London, England; FPO New York 09510-1500		12. REPORT DATE May 1986 13. NUMBER OF PAGES 252
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) Army Research Institute for the Behavioral and Social Sciences. 5001 Eisenhower Avenue Alexandria, Virginia 22333-5600		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

Michael Kaplan, Contracting Officer's Representative

Keywords:

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

parapsychology	ESP	experimenter control
remote viewing	precognition	experimenter bias
random event generator	clairvoyance	
metal bending	probability	
psychokinesis	telepathy	

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report constitutes a critical review of eight major research areas in the field of parapsychology over the past twenty years. The report begins with a philosophical analysis of the way research questions in parapsychology are formulated. It is concluded that the claim to have established psi in the sense of paranormality can be rejected a priori because of the generally conceded absence of a confirmed paranormal theory. Given this fact, the important question becomes whether the observations reported by parapsychologists have

(continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

#20 ABSTRACT- continuation

adequate conventional explanations or whether they are true anomalies. The methods and results of each research project are summarized, along with whatever criticisms of these projects that have been published. This material is then critically evaluated from the point of view of assessing what conventional mechanisms could conceivably account for these findings and the adequacy of these mechanisms as explanations. In general, it is concluded that despite some methodological shortcomings and inadequate reporting, parapsychologists have succeeded in documenting genuine anomalies worthy of scientific interest. Reliable application of whatever paranormal process these anomalies might represent is unlikely until this process (if it exists) is better understood.

FLD 19

ACKNOWLEDGMENTS

I wish to introduce and gratefully acknowledge the important contributions made to this report by the following three individuals, all of whom were supported by the grant.

My Research Assistant, Debra H. Weiner, has been involved in the field of parapsychology for 9 years and has contributed numerous articles to parapsychological publications. In addition to typing and editing the manuscript, her critical comments on the content have been very useful.

I have employed two technical consultants. I chose these particular individuals because (a) they are knowledgable in the areas where I needed technical expertise, (b) they are familiar with parapsychology, (c) they have a level-headed and detached attitude toward parapsychology, and (d) I know them personally and know I can communicate with them effectively.

My statistical consultant was Dr. Donald S. Burdick, Associate Professor of Mathematics and Biomedical Engineering at Duke University. Dr. Burdick has been a consulting editor to the Journal of Parapsychology and is a co-author of a chapter on statistical methods in the Handbook of Parapsychology.

My engineering consultant, whose advice was sought primarily for Chapter 9, was Ronald S. Hawke. Mr. Hawke is a research physicist at Lawrence Livermore National Laboratory. He has a long-standing interest in psychokinesis and is a contributor to The Geller Papers, an edited series of reports on research with Uri Geller.

Whereas the above individuals have given me valuable assistance, all of the chapters in this report have been written by me and I take full responsibility for their content and the conclusions expressed therein.

The research reported in this document has been made possible by Contract number DAJA 45-84-M-0405 from the U. S. Army Research Institute for the Behavioral and Social Sciences through its European Science Coordination Office at the European Research Office of the U. S. Army, London, England. As stated above, the opinions expressed are those of the author; they do not necessarily represent the opinions of the U. S. Army.

TABLE OF CONTENTS

Note: The pages in parentheses refer to the pages in Chapter 10 where the summary of the chapter in question begins.

Chapter 1

INTRODUCTION

The purpose of this report is to critically review experimental research in the field of parapsychology. This introductory chapter has two purposes. The first is to familiarize the reader with the basic terms, methods, and strategies used by parapsychologists in their research, as well as the classes of criticisms leveled against this research by outside commentators. The second purpose, which will require that I indulge in some philosophical analysis, is to propose a reconceptualization of the basic question one must ask in evaluating parapsychological research. The chapter will conclude with a brief discussion of the approach I will take in succeeding chapters to address this basic question.

An Overview of Parapsychology

Parapsychology can be defined as the scientific study of interactions between living organisms and their environment which seem to transcend the currently accepted laws of physics or, more precisely, the so-called "basic limiting principles" of nature, such as those defined by philosopher C. D. Broad (1953):

> (1) General Principles of Causation. It is self-evidently impossible that an event should begin to have any effects before it has happened...
>
> (2) Limitations on the Action of Mind on Matter. It is impossible for an event in a person's mind to produce directly any change in the material world except certain changes in his own brain...
>
> (3) Dependence of Mind on Brain. A necessary, even if not a sufficient, immediate condition of any mental event is an event in the brain of a living body...
>
> (4) Limitations on Ways of Acquiring Knowledge. It is impossible for a person to perceive a physical event or a material thing except by means of sensations which that event or thing produces in his mind...
> (pp. 9-12)

Interactions which transcend these principles are referred to by the generic

term psi.

Psi is traditionally subdivided into two major categories: extrasensory perception (ESP) and psychokinesis (PK). ESP is taken to mean acquisition of information not available to the recognized physical senses or through logical inference. Because of the roots of parapsychology in Cartesian dualism, a further metaphysical distinction is made between telepathy (where the source of the information is assumed to be another mind) and clairvoyance (where the source of the information is assumed to be a material object, event, or process). In those cases where the source of information as such exists in the future rather than in the present, the process is called precognition.

PK refers to the influence of physical objects or events by an organism in ways that cannot be attributed exclusively to known physical forces. During the last decade an analog to precognition has been introduced which postulates PK influence of an event backwards in time; i.e., the effect precedes the cause. This process is referred to as retroactive PK, or retro-PK.

Basic Methodology

ESP. In a test of ESP, the subject is asked to guess a randomly selected target or sequence of targets without access to pertinent sensory information. If another person, called the agent, is attempting to "send" the identity of the targets to the subject, the test is defined operationally as a test of telepathy or of general extrasensory perception (GESP). The latter term is preferred because it takes account of the possibility that the source of information could either be the physical representation of the target or its registration in the mind of the agent. Tests in which there is no agent are referred to as clairvoyance tests. If the targets are not generated until after the guesses are made, it is called a precognition test.

Each attempt to ascertain a target is called a trial, and an uninterrupted series of trials is generally called a run. A correct response on a given trial is referred to as a hit, and an incorrect response as a miss.

The number of hits can be summed over the run or other unit to provide a total score. If the score is higher than the number expected on the basis of chance (called mean chance expectation or MCE), it is called psi-hitting. A below-chance score is called psi-missing. Scores are also sometimes conceptualized in terms of the deviation from MCE irrespective of direction. If the scores deviate widely from MCE the result is called high variance. An overly compressed distribution of scores is called low or tight variance.

Methods for testing ESP can be broken down into restricted-choice (RC) and free-response (FR) categories. In RC tests, the subject is asked to guess a concealed sequence of target items arranged in a random order. The procedure is called restricted-choice because the number of target alternatives and thus the number of scorable responses is fixed and finite.

The traditional targets for RC tests are a deck of cards consisting of five geometric symbols: star, circle, cross, square, wavy lines. A wide variety of standardized test procedures has been developed using these cards (Rhine & Pratt, 1957) and they still see occasional use. A more common procedure, however, is to utilize a device called a random event generator (REG). A random sequence of events is produced through the sampling of an electronic noise source which in some machines is further mediated by the randomly timed emission of beta particles from a decaying radioactive source (Schmidt, 1970b). These decisions are then registered on counters inside the machine or in the memory of a computer to which the device is attached. The subject's task is to identify a symbolic representation of the target state, generally presented to the subject through some sort of display, which the REG has selected (or will select) for each trial. The number of target alternatives generally ranges from two to ten and, much more commonly than with card tests, subjects are given feedback of the identity of the target after each trial. The advantages of REGs over more traditional methods include the more reliable method of randomization and the automated recording of targets and responses, as well as automated counting of correct responses.

In all RC tests, standard statistical techniques are used to determine whether the number of hits or the variance of the scores exceed the theoretically expected value to a significant degree. If this is the case, and to the extent sources of artifact have been eliminated, ESP is claimed to have been demonstrated.

Free-response (FR) tests have become increasingly popular because they are generally more interesting to the subject and more closely approximate the way ESP operates in the "real world." The targets used in FR tests are generally more complex than those used in RC tests. Examples of FR targets are prints of paintings (Ullman, Krippner, & Vaughan, 1973) and View-Master slide reels (Honorton & Harper, 1974). In the highly publicized remote viewing procedure (Targ & Puthoff, 1977), the targets are most commonly geographical sites.

The subject in an FR test is encouraged to free-associate, i.e., to report anything and everything that comes into his mind with the intent that this mentation will pertain to the unknown target. The response period can last anywhere from 5 to 45 minutes, and there is normally just one trial per session. Later, the subject or an outside judge is asked to select on a blind basis from among a set of pictures, sites, etc. (including the target), the one which corresponds most closely to the subject's imagery or mentation report; alternatively, the pictures may be ranked or rated for correspondence on a scale.

These methods ultimately allow the results to be evaluated statistically in ways comparable to those used for RC tests. However, this is accomplished at the price of a great loss in power such that statistical significance can rarely be demonstrated for a single session. Some more powerful techniques which involve breaking down the targets and/or responses into discrete information units have occasionally been applied (e.g., Jahn, Dunne, & Jahn, 1980).

The subject is often prepared for a FR test through induction of a hypnagogic-like state of consciousness designed to break down linear thought processes, encourage inward focusing of attention, facilitate the flow of mental imagery, and eliminate distracting external stimulation. The most popular of such techniques is the ganzfeld, a procedure in which the subject looks through halves of ping-pong balls covering the eyes into a white or red light while listening to white or pink noise being played through headphones (Bertini, Lewis, & Witkin, 1969). This procedure often produces effects somewhat similar to longer-term perceptual deprivation, but without the adverse side effects.

PK. The traditional method of PK testing utilizes mechanically thrown dice, the subject's task being either to make one face appear uppermost or to cause the dice to fall on one side or the other of a divided surface (Rhine & Pratt, 1957). However, dice tests have not been used for many years. By far the most common method of contemporary PK testing is to have the subject attempt to bias the output of an REG by influencing the electronic noise or radioactive decay processes. If desired, trials can be generated at very rapid rates (hundreds per second), which allows for the application of powerful statistical analyses. Ongoing analog or digital feedback can be provided to subjects in innumerable ways in either the visual or auditory mode, and the feedback display itself is often presented to the subject as the target (e.g., "Keep the red line above the center of the screen"). Methods of statistical analysis are comparable to those employed in ESP tests.

A wide variety of other techniques involving an equally wide range of physical processes have seen limited use. Those which seem most likely to evolve into standardized experimental paradigms involve the subject attempting to produce localized changes in temperature as measured by thermistors (Schmeidler, 1973) or stress in metalic objects as measured by strain gauges or piezoelectric sensors (Hasted, 1981). The more highly publicized gross

metal-bending procedures in which the subject is allowed physical contact with the target specimen are notoriously difficult to control and generally not of scientific interest unless anomalous molecular transformations in the structure of the specimen can be demonstrated.

One finds even less standardization in the study of PK effects on biological systems. Examples of target systems in past research range from digestive enzymes in vitro (Smith, 1972) to bacteria (Rauscher & Rubik, 1980) to skin lesions in mice (Grad, Cadoret, & Paul, 1961). Serious "healing" research with humans is virtually non-existent, but some attempts have been made to remotely influence psychophysiological responses such as GSR (e.g., Braud, 1978).

Research Strategies

There are two major research strategies which parapsychologists have adopted. Proof-oriented experiments, which to the extent they are limited to this strategy could more properly be called demonstrations, involve attempts to demonstrate psi effects in such a way that all reasonable "normal" or conventional explanations have been ruled out. Almost all psi experiments which are widely known outside of parapsychological circles are primarily or exclusively proof-oriented.

The majority of psi experiments, however, are primarily process-oriented. In its pure form, this approach avoids tackling the ontological status of psi directly and attempts instead to identify its psychological and physical correlates as a basis for the development of explanatory theories or models. Psi scores are treated as dependent variables to be related to such things as scores on psychological tests and manipulations of physical or psychological conditions as independent variables. Because of the need to improve the reliability of psi effects, particular interest has been directed toward identifying test conditions that are psi-conducive. Ultimately, however, the value of this approach will be determined by its capacity to

develop networks of reliable correlates of psi effects that differ from what might be predicted on the basis of conventional counterexplanations of these effects.

Most process-oriented research is guided by implicit theory and sometimes by more fully developed models from which testable hypotheses are derived. Most theorizing in parapsychology is psychologically oriented and addresses such issues as how ESP "information" is processed, blocked, or distorted after it reaches the mind of the subject. The most fully articulated and comprehensive of these psychological models is that of Psi-Mediated Instrumental Response (PMIR), which links both ESP and PK to principles of learning theory and dynamic psychology (Stanford, 1977).

Theorizing about how information gets from the source to the receiver in ESP, or how the subject affects the target system in PK, draws more heavily on physics. The most fully developed specimens here are the so-called Observational Theories (OTs), especially the version of Walker (1975). These theories represent extensions or radical interpretations of quantum mechanics, their main premise being that observation of the data of a psi experiment serves a function analogous to measurement in quantum mechanics. The notion of retro-PK is a direct consequence of these theories. The OTs have generated some testable predictions as well as much controversy.

Most process-oriented psi experiments are also proof-oriented in the sense that attempts are made to incorporate the kinds of controls demanded of proof-oriented experiments. Nonetheless, the objectives of the two kinds of experiments are clearly different.

Criticisms

External critics of parapsychology generally have not acknowledged the existence of the process-oriented approach in psi experimentation, so their criticisms are directly relevant only to the proof-oriented approach. The major points of attack can be classified under the following headings:

Fraud. Because of the origins of parapsychology in Spiritualism and the fact that a high percentage of its external critics are either amateur or professional stage magicians, suggestions of fraud by high-scoring subjects have become commonplace. Some critics also consider it appropriate to speculate about fraud on the part of experimenters. Isolated cases of experimenter fraud have in fact been uncovered in parapsychology (e.g., Rhine, 1974), but the extent to which such transgressions can be generalized to the field as a whole is debatable.

Sensory Cues. In ESP experiments, the subject should have no access to sensory information about the target. Critics are not always satisfied that such cues have been eliminated. Hyman (1985), for example, has noted that in some FR experiments the target picture handled by the agent is included in the set of pictures later given to the judges for scoring and could contain identifying fingerprints, etc.

Randomization. It is generally considered to be important that the target sequences in ESP experiments be satisfactorily random. This is particularly crucial in those experiments where subjects are given trial-by-trial feedback of targets and could learn to identify patterns in the sequence during the course of the test. Likewise, in REG PK experiments it is considered important that the output of the REG be satisfactorily random in the absence of attempted PK influence. Whether adequate procedures have been used both to generate and to verify randomness has been a major focus of criticism of psi research.

An alternative to establishing randomness, which is necessary in those PK procedures where theoretical chance baselines cannot be defined, is to compare psi test results to empirically defined baselines established in control conditions.

Statistics. Statistical criticisms of psi experiments are difficult to compartmentalize. Nonetheless, it would be fair to say that most concern either alleged violations of the independence assumptions of the statistical

test employed or failure to adjust significance levels for multiple analyses.

Data Selection. It is sometimes suggested that parapsychologists withhold nonsignificant results from their reports or classify them as exploratory on a post-hoc basis, thereby making the reported results seem more significant than they really are. A related criticism is labeled optional stopping, which refers to aborting an experiment or series of trials at a randomly occurring apex in the scoring level. It is generally agreed that it is permissible to apply optional stopping to subunits of trials (e.g., the number of trials contributed by a particular subject in a multi-subject experiment) so long as the total number of trials in the experiment is specified in advance.

Replicability. Although replicability by itself cannot establish the paranormal nature of psi anomalies, most parapsychologists and their critics agree that it is a necessary prerequisite for establishing the reality status of psi. No one claims that psi effects are reproducible on demand, but many parapsychologists claim that certain psi effects are replicable to a degree that significantly exceeds chance expectancy, i.e., statistical replicability. Various attempts have been made to demonstrate this claim through a technique that has come to be called meta-analysis (Glass, McGaw, & Smith, 1981), in which groups of experiments are treated statistically in much the same way as are groups of subjects in individual experiments.

A closely related problem is that successful psi experiments are not randomly distributed among the investigators who conduct them. In other words, while some experimenters in parapsychology seem to consistently obtain significant evidence of psi in their experiments irrespective of the particular type of experiment undertaken, others just as consistently do not. This so-called experimenter effect looms as a major problem in the field and has obvious implications for the replicability issue. The fact that successful experimenters tend to be those favorably inclined to the reality of psi has led many critics to conclude that such experimenters are successful because

their "belief" in psi causes them to overlook potential artifacts. Parapsychologists, on the other hand, have suggested that successful experimenters either are better at motivating their subjects or are the sources of the psi themselves. Some empirical evidence has been offered in support of both of these hypotheses.

Asking the Right Question

In Search of the Conclusive Experiment

We have now come to the point where it is necessary to examine how parapsychologists have formulated their basic research objectives. Parapsychological inquiry has traditionally organized itself around the question, "Does psi exist?" The term psi, as noted in the preceding section, is defined negatively as some process that transcends currently accepted physical principles. It is not surprising, therefore, that the approach to its verification or validation has also been negative. Again as previously noted, psi is considered to have been demonstrated if, and only if, all conventional processes, i.e., processes subsumed under the basic limiting principles, have been eliminated. Both parapsychologists and their critics have agreed on this requirement. Indeed, the controversy around the pioneering experiments of J.B. Rhine in the 1930s focused on just this question: Did any of Rhine's experiments in fact eliminate all such possibilities?

Rhine, perhaps influenced by the simplistic behaviorism which reigned in psychology at the time, overestimated the ease with which this requirement could be met. In Extrasensory Perception After Sixty Years (Rhine, Pratt, Stuart, Smith, & Greenwood, 1940) he and his colleagues painstakingly analyzed all the experimental work up to that time with reference to 35 conventional mechanisms proposed by critics, which included faulty statistics, data selection, matching biases in target and response sequences, shuffling defects, recording errors, sensory cues, and experimenter incompetence. Six

experiments were found to be immune to all 35 criticisms, of which the two most prominent ones were the Pearce-Pratt Series and the Pratt-Woodruff Series. However, critics, most notably C. E. M. Hansel (1966), had little difficulty in pointing out ways in which the statistically significant results of these two experiments could be explained by conventional processes: in the case of the Pearce-Pratt experiment, it was by cheating on the part of the subject; in the case of the Pratt-Woodruff experiment it was by cheating on the part of the junior experimenter. Rather indignant exchanges about both experiments raged in the literature into the 1970s, with no clear resolution.

With the benefit of 20/20 hindsight, it is nonetheless fair to say that both experiments could have been better designed to take account of the possibilities raised by the critics. It is equally clear that no other psi experiment has been shown to be immune to all conceivable counterexplanations. In fact, parapsychologists no longer claim such an experiment. The approach nowadays is to argue either that the "flaws" cited by critics in response to the better experiments are trivial and speculative (i.e., the counterexplanations are implausible) or that the collective weight of the experiments is compelling even though no single experiment by itself is conclusive.

On the other hand, parapsychologists have been reluctant to repudiate explicitly the proposition that an evidential psi experiment must eliminate all conventional alternatives, probably out of the quite reasonable fear that to do so would expose them to charges of sloppiness, lowering methodological standards, etc. This reluctance has allowed critics to argue persuasively that parapsychologists have failed to establish the existence of psi by their own (parapsychologists') criteria.

However, the fact remains that the standard of the conclusive experiment is encumbered by logical difficulties which are both real and fatal. Criticisms such as experimenter fraud, if carried to their logical conclusion, are patently unfalsifiable, as even some of Hansel's fellow critics recognize

(e.g., Hyman, 1981). Although Hansel replies that sufficient independent replication of psi experiments would suffice to overcome such criticisms, this conclusion does not follow logically from his premises. The replicability of an effect, however consistent that might be, tells us nothing about its causal mechanism. Even collusion or fraud on the part of all the experimenters, although a rather implausible scenario, would be a preferable explanation to psi, according to the logic of Hansel's position. In fact, Hansel is rather explicit in stating that the implausibility of a conventional hypothesis should not be held against it: "A possible explanation other than extrasensory perception, provided it involves only well-established processes, should not be rejected on the grounds of its complexity." (Hansel, 1980, p. 21)

But even if a critic were to concede the honesty of the experimenter (or, for that matter, the subjects) and no other counterhypothesis could be put forth, it still would not follow that all such counterhypotheses have been ruled out. The reason is simply that one cannot be sure that all counterhypotheses have been thought of at a given point in time. It is therefore legitimate, as Hyman (1981) has in fact done in relation to the successful PK experiments of Helmut Schmidt, to ask that we suspend judgment for an unspecified period of time, banking on the idea that an acceptable counterexplanation will eventually emerge. The problem, however, is that the possibility of conventional counterexplanations can never be ruled out because the population of such counterexplanations can never be defined in a way that is known to be adequate. In other words, since one can never know if all possible counterexplanations have been thought of, one must suspend judgment indefinitely.

The implication of the preceding analysis is simply that the presence or absence of a "conclusive" experiment, even a repeatable one, is not an adequate standard by which to evaluate the claim "psi exists," because it is inherently unfalsifiable. So what is an adequate standard?

Conflation in the Use of "Psi"

Before answering this question, it is necessary to consider a curious characteristic of the way the term "psi" is used both by parapsychologists and their critics. According to the official definition, "psi" refers to a paranormal principle or cause; i.e., it is intended to be a theoretical or at least a quasi-theoretical construct that serves to explain certain natural events. However, the term "psi" is often used as well to label the events themselves, as in the more complex term, "psi phenomena." The point here is that with respect to actual usage, no clear distinction is made between the phenomena under study and the quasi-theoretical principle proposed to account for them, between the explanandum and the explanans.

One illustration of this conflation is the accepted definition of parapsychology: "the scientific study of paranormal phenomena" (Thalbourne, 1982, p. 51), which can be translated as "the scientific study of psi." Note that the definition assumes that the paranormality of the phenomena under investigation is granted a priori. This of course does not adequately describe most parapsychological research, which does not assume paranormality a priori but rather is undertaken to verify paranormality a posteriori, empirically. The definition, however, defines the subject matter of parapsychology in terms of parapsychologists' preferred explanatory framework.

The same conflation can be detected in the writings of critics when they claim that parapsychology lacks "facts" or a subject matter. What they really mean is that parapsychologists have failed to establish the "existence of psi." However, what parapsychologists have failed to establish is psi the theoretical principle, i.e., psi the explanans. But a theoretical principle is not a subject matter. The subject matter of parapsychology is its phenomena, the explanandum. Only if we conflate the explanandum and the explanans does the statement that parapsychology lacks a subject matter seem to make sense.

The way out of the conflation is simply to define the phenomena parapsychologists study in a theoretically neutral way, that is, independently of whether the phenomena are in fact paranormal. In a previous paper (Palmer, 1985) I have suggested that the term psi be retained to label the phenomena and proposed the term omega to label the theoretical or quasi-theoretical principle proposed by parapsychologists to account for them. Only omega implies paranormality. While we await the adoption of this or some comparable scheme, I have suggested that "psi phenomena" be labeled as "ostensible psychic events" (OPEs).

Rephrasing the Question

Appreciation of this conflation encourages a critical examination of how the fundamental research problem in parapsychology is phrased. The existential phrasing of the question "Does psi (i.e., paranormality) exist?" both reflects and reinforces the conflation of explanandum and explanans because "existence" is more naturally attributed to the former than to the latter. Indeed, reification of a theoretical construct is often considered objectionable in the philosophy of science. In any event, the preceding analysis suggests a better phrasing of parapsychology's fundamental research question: "How can ostensible psychic events (OPEs) be best explained?"

This new question has several important implications which bear upon our original question of what is the appropriate standard for evaluating evidence for psi. One is that parapsychologists can only "demonstrate" paranormality by confirming a theory that adequately explains OPEs by appeal to some "paranormal" theoretical principle, i.e., a theoretical principle that transcends Broad's basic limiting principles. This means that paranormality would not be established even if a conclusive experiment were both possible and replicable on demand. "Paranormality" can only be legitimately claimed in the context of a theory that positively defines some paranormal principle. In other words, the parapsychologists' attempt to "demonstrate" paranormality by

eliminating competing alternatives is logically flawed and can be rejected prior to analysis of their success in actually eliminating the alternatives considered.

Have parapsychologists succeeded in establishing paranormality by the more proper route, i.e., by confirming a paranormal principle or theory? Even the great majority of parapsychologists would concede that this has not yet been accomplished. Although the Observational Theories represent a serious attempt in this direction, these theories have not yet been sufficiently tested to be considered established.

On the other hand, and this is a key point, the failure of the parapsychologists to provide an adequately verified paranormal explanation of OPEs does not imply the existence of adequate conventional explanations. Another of the unfortunate consequences of the question "Does psi exist?" is that it has caused parapsychologists and critics alike to assert that the burden of proof in parapsychology falls exclusively on the claim of paranormality, i.e., the claim that "psi exists." The main rationale for this conclusion is that it is unreasonable to demand verification of the opposite conclusion, "Psi does not exist," because it is a universal (and existential) negative. But this is no longer the case when the question becomes "How can OPEs be best explained?" Here the canons of scientific method clearly state that the burden of proof falls upon anyone who proposes to explain OPEs, whether the appeal be to paranormal or conventional explanations.

OPEs for which no adequate explanations have yet been found can be construed as anomalies with respect to the basic limiting principles, because when taken at face value they are inconsistent with them (Palmer, 1985). Calling them anomalies is in no way meant to imply that the explanations of OPEs are necessarily paranormal, or that an adequate conventional explanation of OPEs may not someday be found. However, the fact that such events are paranormal when taken at face value is considered reasonable grounds for including paranormal explanations among the population of possible or

potential explanations of OPEs that need to be considered.

Although parapsychology experiments are routinely construed both by parapsychologists and their critics as tests of the "psi hypothesis," the present analysis suggests that in most cases they could more profitably be construed as tests of certain specific conventional hypotheses which purport to explain OPEs. Only rarely do such experiments test a paranormal theory or mechanism. Thus, the real issue in most experiments is not whether OPEs are paranormal but whether they are anomalous. The results of such experiments are anomalous to the extent it can be shown that no conventional explanation of the results is scientifically adequate.

Redefining the Standards of Evidence

The most difficult question confronting this analysis is what criteria should be set for an adequate scientific explanation of OPEs. Some would argue on philosophical grounds that one such criterion is that the explanation must be conventional, based on appeal to the so-called "coherence" principle. This principle states that the currently accepted laws of nature, which preclude paranormal processes, are universal in scope. Although the coherence principle has not always been a reliable guide in science, Newtonian mechanics being its most notorious failure, it is nonetheless positively valued in the scientific community and I cannot logically compel its abandonment. On the other hand, no empirical evaluation of parapsychological research, such as will be attempted in this review, would make sense if the coherence principle were to be accepted in its strongest form. It is worth noting that a moderately strong form of the coherence principle plays a prominent role in the approach of most critics of parapsychology, especially those like Hansel who argue that all conventional hypotheses must be ruled out before paranormal hypotheses can be entertained.

The remaining criteria for a scientifically adequate explanation of OPEs, apart from logical (internal) coherence and the like, are the same empirical

criteria we see in the rest of science. Whereas these are hard to define on paper and disagreements abound as to how well they are met in specific instances, at least the problem we are dealing with is a familiar one.

It is perhaps worth mentioning in passing that many scientists (including some parapsychologists) who accept a weak form of the coherence principle would argue that a greater degree of empirical evidence is necessary to support a paranormal theory than a conventional one, that "exceptional claims require exceptional proof." I admit to being somewhat of a maverick in rejecting this proposition. Briefly, my reasons are the following: (1) applying such a principle leads to selective rejection of research findings and a bias in the research literature that would artifactually favor a conventional theory; (2) a conventional theory that really works should not need such a crutch; and (3) in the case of OPEs, confirmation of a paranormal theory would not logically require abandonment of any conventional theory but simply a redefinition of its boundaries. My own position is that standards of evidence should be uniform (and rigorous) throughout science. However, this issue is not, strictly speaking, relevant to the present review since paranormal and conventional theories are not being contrasted; for the most part conventional hypotheses are being examined in isolation.

The history of parapsychological criticism clearly shows that it is easy to devise ad hoc conventional explanations of the OPEs that appear in laboratory experiments. However, a <u>possible</u> explanation is not the same as a <u>scientifically adequate</u> explanation. But how is it possible in practice to assess the scientific adequacy of conventional explanations of the results of particular psi experiments?

I will propose the following three guidelines:

(1) <u>Internal empirical evidence within the experiment itself</u>. Sometimes the conventional hypothesis leads to predictions that can be tested by new analyses of the data from the experiment under consideration.

(2) Empirical support for the hypothesis in related contexts. This might include confirmation of the hypothesis in related experiments or experiments explicitly designed to test the hypothesis.

(3) Plausibility. This is a hard term to define and is admittedly subjective. In a nutshell, it is simply a commonsense judgment of the likelihood that a conventional process would take place. It might include such things as the difficulty or complexity of the process, the apparent motivation of a subject to undertake it (as in the case of fraud), etc.

Perhaps the best summary guideline might be the following: Would we be willing to accept a particular conventional hypothesis if the experiment were an "ordinary" one and the controversial question of paranormality were not involved? Often there is a temptation to accept a conventional hypothesis simply because the alternative (paranormality) is seen as intolerable. The preceding question helps us to avoid this temptation.

General Approach

In the remainder of this report, I will explore the question of whether experimental data exist which can be properly classified as anomalous, data for which the available conventional explanations are inadequate (even if possible) and the possibility of paranormal causes must, therefore, be seriously considered. A great deal of research relevant to this question has been published in parapsychological journals over the last century. Two approaches can be taken to reviewing this material. The first is to provide an overview of the entire literature, and the second is to provide a more in-depth review of the most potentially evidential subsections of this literature. I have chosen the second approach for two reasons. Although the first approach can serve useful functions, particularly for those sympathetic to the concept of paranormality who are looking for promising hypotheses for process-oriented research, it is not likely to be very satisfying to the critically oriented reader for whom this report is intended. Detailed

analysis of specific research programs is simply not possible within the framework of a broadly based review. My second, more pragmatic reason for choosing the second approach is that I and others have already written relatively current reviews of the first type (see Krippner, 1977, 1978, 1982, 1984; Wolman, 1977).

Commitment to the second approach raises the question of which research programs should be selected for review. For the most part, I have avoided trying to apply some objective formula but have relied instead upon my own professional judgment, based on 15 years of experience in the field of parapsychology, in making my selections. Nonetheless, there are certain general principles which guided my thinking. These include the following:

(1) The research must represent an integrated body of experiments using a similar methodology. "One-shot" studies, however impressive, were not considered unless they could be related to similar studies by other investigators.

(2) On the surface, the research program must have yielded statistically significant results with at least moderate consistency.

(3) The research program must be considered important and evidential by a significant proportion of contemporary parapsychologists and, preferably, achieved sufficient notoriety to evoke responses by outside critics. (An exception was made on this point for the research on metal bending. Even though this research is not highly regarded by most parapsychologists, it represents an important new research direction with potentially far-reaching implications.)

I have chosen to evaluate eight classes of parapsychological research programs which have been conducted since 1970. Each of the following chapters (2-9) is devoted to one of these classes, and several of the chapters review more than one program. Eight major research programs conducted by a particular parapsychological investigator or research team are reviewed. The principal investigators and their affiliations at the time the research was

conducted are as follows: John Bisaha and Brenda Dunne (Mundelein College); Charles Crussard (Pechiney-Ugine-Kuhlmann Aluminum Company, Paris); John Hasted (University of London); Robert Jahn (Princeton University); B. K. Kanthamani and Edward Kelly (Institute for Parapsychology, Durham, NC); Harold Puthoff and Russell Targ (SRI International); Rex Stanford (St. John's University); and Montague Ullman and Stanley Krippner (Maimonides Medical Center). In addition to the above, two chapters (4 and 7) are devoted to groups of experiments on common themes conducted by a wider range of investigators. Summaries of each of these chapters are presented in Chapter 10. The reader may find it helpful to peruse the summary of a given chapter before turning to the chapter itself.

Each of the chapters 2 through 9 is organized in more or less the following manner:

(1) A description of the methodology employed in the experiments;

(2) A description of the results obtained and their interpretation by the investigators;

(3) A description of published criticisms of the research;

(4) My own evaluation of the research and the criticisms.

A few additional comments on the last component are in order at this point. First, the reader has a right to know something about my own background and involvement with the field of parapsychology. My training is as an experimental psychologist, with my specialty in the area of personality/social psychology. As noted previously, I have been involved in parapsychological research for 15 years, and I thus could be considered an "insider." Parapsychologists are an extraordinarily close-knit group, and I am thus on a first-name basis with the great majority of the parapsychologists (as well as several of the critics) whose work I will be reviewing. I do not feel that this fact has compromised by objectivity, and in at least two cases I have introduced novel criticisms of research conducted by investigators whom I consider to be personal friends.

My involvement in the field has obviously conditioned the attitudes I have brought to this task. On this score, I view myself, and I think I am viewed by most of my colleagues, as a moderate. On the one hand, I would not have remained in parapsychology this long if I did not feel that there was "something to it," that the field is potentially very important. On the other hand, I am impressed with how little we know about the causes which underlie the effects we study in parapsychology, and I tend to react negatively to "extremists" on both sides who make claims or draw conclusions that in my opinion outstrip the evidence.

Given the above, the reader should not be surprised to discover that I will not be drawing definitive conclusions about the evidence reviewed in this report. How one evaluates the evidence inevitably comes down to the plausibility one attaches to the "normal" explanations which can be attached, just as inevitably, to any piece of psi research. The question the reader must constantly ask himself or herself in the following pages is how far the researchers have succeeded in pushing these "normal" explanations in the direction of absurdity. These judgments will inevitably involve a subjective component, and reasonable people can be expected to differ in the judgments they make. The best I can do as a reviewer is to point out what the known "normal" explanations are and what must be taken into account in assessing their plausibility. Although I feel responsibility as a reviewer to express my own opinions about their plausibility, I also encourage readers to feel free to draw their own conclusions.

Chapter 2

THE MAIMONIDES DREAM EXPERIMENTS

The first major ESP research project in the modern era to use free-response methodology was a series of experiments conducted at Maimonides Medical Center in Brooklyn, New York, exploring telepathy in dreams. The principal investigators were psychiatrist Montague Ullman and psychologist Stanley Krippner, with major contributions also being made by Charles Honorton.

The basic method was to have an agent attempt to influence the dreams of a percipient by concentrating on a randomly selected art print. Later the percipient(s) and/or outside judges would attempt to match up the targets for the series with the dream protocols on a blind basis, using standard methodologies for judging free-response ESP materials. Generally, only one trial was collected per night.

The Maimonides experiments can be divided into three categories:

(1) Formal Experiments: One Trial per Subject. This category includes two screening experiments in each of which twelve paid volunteers participated as subjects (Ullman, Krippner, & Feldstein, 1969; Ullman & Krippner, 1970). I have also included in this category one other experiment in which selection criteria were somewhat more rigid, i.e., subjects were to have reported spontaneous telepathic experiences or to be acquainted with the agent (Krippner, Honorton, Ullman, Masters, & Houston, 1971).

(2) Formal Experiments: Multiple Trials per Subject. In these experiments, subjects selected either on the basis of promising results in the screening experiments or because for other reasons they were expected to perform well in this type of task participated in four to eight sessions each. Two graduates of the first screening study were invited back: a psychiatrist,

William Erwin, was the subject in two experiments consisting of seven and eight trials, respectively (Ullman et al., 1969; Ullman & Krippner, 1970), and a secretary, Theresa Grayeb, who completed one eight-session experiment (Ullman & Krippner, 1970). One graduate from the second screening experiment, a psychologist named Robyn Posin, completed an eight-session experiment (Ullman & Krippner, 1970). The remaining subjects, who had not participated in the screenings, were psychologist and parapsychologist Robert Van de Castle and a psychic named Malcolm Bessent. Van de Castle was the subject for one eight-night series (Krippner & Ullman, 1970). Bessent was the subject for two eight-night series using a precognition procedure (Krippner, Ullman, & Honorton, 1971; Krippner, Honorton, & Ullman, 1972), and one four-trial telepathy series in which the agents were the audience of a rock concert (Krippner, Honorton, & Ullman, 1973). Another psychic, Felicia Parise, served as a control percipient in this experiment; i.e., the audience was unaware of her involvement. This group of experiments was obviously the most important in the project because it was restricted to subjects who were expected to succeed.

(3) Informal Pilot Sessions. Several hundred pilot sessions were conducted during the course of the research project and reported in unpublished manuscripts. The methodology was the same as that of the formal experiments with respect to basic controls.

Methodology

Targets and Target Selection. The targets for the Maimonides experiments were usually postcard-sized prints of famous paintings selected for simplicity and distinctiveness of detail and, in later series, emotional evocativeness. Also in later series, the prints (or slides) were supplemented with multi-sensory materials to increase the salience of the targets. These varied from toy objects in the second Erwin experiment to

appropriate recorded music in one of the group experiments. This latter experiment will subsequently be labeled as the "sensory bombardment" experiment (Krippner, et al., 1971).

Sets of targets were assembled for each experiment, each set generally equal to the number of trials in the experiment. The prints in each set were selected to be maximally diverse in content. In the early experiments the target pools were selected by the agent and experimenter but later on this task was performed by a third party not involved with the actual conduct of the sessions.

The agent selected the target (without replacement) from the prints remaining in the pool. Procedures varied somewhat from experiment to experiment, but in all cases except possibly one (Krippner, et al., 1973) the target was determined by a digit from a random number table, the designation of the digit in turn being determined by a complex quasi-random procedure. Some of these selection methods are problematic and will be discussed further in the evaluation section.

Test Procedure. Again, the procedures for the test sessions varied slightly from experiment to experiment, but the following account is representative.

When the percipient arrived for the session, he or she was allowed to meet with the agent to establish rapport. The agent was a member of the lab staff and in some studies the percipient was given some choice in determining the agent for a given session. The percipient then got ready for bed and electrodes which measure EEG and eye movements were applied. During the course of the night the pattern of brainwaves and eye movements were monitored by the experimenter, located in an adjacent room, to determine those times at which the percipient was likely to be having a dream dominated by visual imagery (as opposed to verbal mentation). These periods are called "rapid eye movement" or REM periods and occur about three

to six times per night.

Once the subject was in bed, the agent went to a room at the other end of the building and selected the target picture. Periodically during the night the agent attempted to "send" the contents of the target to the percipient. In later studies, the experimenter signaled the agent by a buzzer, indicating the onset of a REM period, so that the sending could be yoked to the percipient's dreams. Toward the estimated end of each REM period, the experimenter awakened the subject and elicited a dream report, which was taped.

In the morning, the experimenter played back the tapes of the dream reports and asked the percipient to add any associations he or she might have had to the dream mentation and to venture a guess as to the identity of the target. These associations were also taped. Collectively, this material constituted the dream protocol for the session.

The intercom set-up allowed no communication from the agent's room to either the percipient's room or the experimenter's room. The agent had no contact with the percipient until after the session and percipient judging (if this was done) was finished.

The possibility of sensory cues was further minimized in the two precognition experiments with Bessent. In these experiments the "agent" selected the target for the night and displayed it to the percipient in the morning after the dream protocol had been completed.

Judging. In most cases judging was undertaken both by the subject and by outside judges (usually three) who worked independently of each other. (In several cases, one or more other judges conducted supplementary judgings.) At the end of an experiment, which consisted of from four to twelve sessions, each judge was asked to rate each possible target-transcript pair on a 100-point scale indicating confidence in it being a correct match. These ratings are thus essentially a refinement of

rankings. Judges also evaluated the dream protocols both with and without the morning-after associations. In some cases, ratings were also based on the subject's "guess for the night," an assessment based upon his dream reports and associations.

Most of the reports contain no information about the order in which the targets and protocols were given to the judges. However, in three of the experiments (the second Erwin experiment and the two precognition experiments) rankings were not used and the judges were asked to rate all possible target-protocol pairs in random order.

In the experiments in which subjects completed only one trial, the subject ranked and rated his or her protocol against each of the potential targets in the experiment at the end of the session. This only applied to the screening sessions. In the experiments with multiple trials per subject, the subject performed the same judging task as the independent judges after all sessions had been completed. However, subject judging was not used in the "sensory bombardment" experiment, the second Erwin experiment, or the precognition experiments.

Judging by both subjects and independent judges was always done blind and duplicate target sets were always used; i.e., the print handled by the agent was never included in the judging material.

Statistical Analysis. A variety of methods of analysis were employed and multiple methods were frequently used in the same experiment. Regarding the ranks, hits were defined either as a rank of one (direct hit) or, more commonly, as a rank in the lower half of possible ranks (binary hit). Significance was then determined by a simple binomial or exact probability test. Ratings were evaluated by comparing the mean rating (averaged over the outside judges) assigned to the correct target-transcript pairs to the mean rating assigned to the incorrect pairs using one of a variety of methods varying from Scheffé analysis of variance (Scheffé, 1959, Ch. 10)

to a latin square ANOVA to a Mann-Whitney U test.

In summary, the analysis options were distributed along the following three dimensions:

1) Independent judging and subject judging;

2) Dream protocols with and without morning-after associations, and "guess for the night;"

3) Rankings and ratings.

Thus, several tests of the hypotheses were customarily included in the reports.

Results

It is sometimes but by no means always clear which analysis or analyses had been designated in advance to be the primary test of the hypothesis. Fortunately, in most cases the analyses converged on a common conclusion.

Formal Experiments: One Trial per Subject. The two screening experiments both yielded nonsignificant results. However, in the first screening experiment, post-hoc analysis revealed that the results of those subjects tested when the male research assistant served as agent and the female as experimenter were significantly positive and significantly better than those when the roles were reversed. Results from the Krippner et al. (1971) study were significantly positive for independent judging but not for subject judging.

Combined, these three experiments produced 21 binary hits from 32 trials (66%) based on the rankings (or converted ratings) of the independent judges as applied to the total transcripts (dreams plus associations). This is associated with a corrected Z (Z_c) of 1.59, which is not significant.

Formal Experiments: Multiple Trials per Subject. The two experiments with Erwin, the one with Van de Castle, and the three with Bessent all

yielded significant positive results. The single experiment with Grayeb and the single experiment with Posin yielded chance results. The results of Parise, the control subject in one of the Bessent experiments, were also close to chance.

Combined, these experiments produced 49 binary hits from 67 trials (73%) based on the rankings (or converted ratings) of the independent judges. This is associated with a Z_c of 3.67, which is significant at $p<.001$.[1] (The Parise results are included since it can be argued that agents' focus of attention on the percipient may not be necessary for psi to occur in this paradigm.)

Pilot Experiments. Of the 280 pilot trials evaluated by independent judges, 165 were binary hits (59%). This is a smaller percentage than was found with the other single-trial-per-subject experiments, but due to the larger sample size it is significant ($Z=2.99$, $p<.01$).

The probability values reported above do not take into account the multiple analyses employed by the authors or possible dependencies in the judgings and thus should be considered approximate. Additional analyses will be presented in the evaluation section. Nonetheless, these analyses, along with the fact that seven of the eleven formal experiments were significant (six of eight with selected subjects), suggests that, taken at face value, the research project as a whole yielded results exceeding chance expectancy.

Wyoming Replications

Single replications of two of the successful Maimonides experiments, the Van de Castle experiment and the "sensory bombardment" experiment, were undertaken by dream researcher David Foulkes and colleagues at the University of Wyoming (Belvedere & Foulkes, 1971; Foulkes et al., 1972). Both experiments were designed in consultation with the Maimonides team.

Van de Castle served as the subject in the first replication, and subjects in the second replication were selected on the basis of the same criteria (spontaneous telepathic experiences and rapport with the agent) as in the original experiment. In the sensory bombardment replication, a main difference was that, in contrast to the original experiment, the agent (in New York) was several thousand miles away from the percipient (in Wyoming). However, since distance does not appear to be a critical limitation to ESP, this modification was considered acceptable by all parties concerned.

The experimental procedures of the replications closely followed those of the original studies. The most notable differences were that the targets for each night were selected by an additional experimenter in the Wyoming experiments whereas they had been selected by the agent in the Maimonides experiments. Also, the agent could not leave his or her room in the Wyoming replication of the Van de Castle study. (The door and windows were sealed shut.) Such elaborate precautions were not taken in the Maimonides experiment.

Judging was performed by both the subject and two independent judges in the Van de Castle replication and by three independent judges in the "sensory bombardment" replication. Only rankings were used. The results were nonsignificant for both experiments.

Criticisms

The most extensive criticism of the Maimonides experiments has been offered by the British psychologist C.E.M. Hansel (1980) who for many years has been the most prolific critic of major psi experiments. His critique of the Maimonides experiments dwelled exclusively on the possibility of sensory leakage in the Van de Castle experiment, which he compared unfavorably to the replication attempt by Foulkes in this respect. His main specific point was that in the experimental report which he used, the description of the method implies that "an experimenter appears to have been with the agent

when he opened his target envelope" (p. 246). This of course would mean that the experimenter, who elicited the dream reports from the subject, was not blind to the target.

Another criticism, made primarily by psychologist James Alcock (1981), is that there was no control judging to provide an empirical baseline. This would require that the targets in the control judging be assigned in a random order. He acknowledged that the Maimonides team did perform such a control judging for one of the successful experiments (the second Erwin experiment) but he considered this inadequate.

Psychologists Leonard Zusne and Warren Jones (1982) suggested that in some of the experiments the percipient was shown the target prior to collection of the dream reports. This is a misunderstanding of the procedure which perhaps reflects the fact that they used as their source a brief description of the second Bessent precognition experiment which appeared in a popular book (Ullman & Krippner, 1978). In this particular experiment, sessions designed to test for precognition were alternated with other sessions designed to determine whether the experience of observing the precognition target for the night before would affect dream mentation during the night following. This brief description of the procedure apparently left Zusne and Jones with the impression that these latter sessions were meant to be the precognition sessions.

Finally, psychologist Irvin Child (in press) pointed out that in most of the series in which a subject completed multiple trials it cannot be assumed that the judgings were independent as required by the statistical tests employed. Although judges were instructed to assess the trials independently, it cannot be assumed that this independence was achieved in practice. The only experiment of this type to which this criticism is inapplicable is the Van de Castle experiment where a separate target <u>pool</u> was used for each session.

Child, however, attempted to show that this criticism is not fatal by demonstrating that the results of judgings to which the criticism does not apply (including some judgings in the single-trial-per-subject experiments and all judgings from the pilot sessions) were collectively significant. This was accomplished by taking the most sensitive analysis available from the reports, converting the result to a Z, and combining the Zs by the Stouffer method (Mosteller & Bush, 1954). The resulting p-values were less than .002 for subject judging and less than 10^{-6} for independent judging.

Evaluation

Statistical Independence

Child's criticism of the statistical methods employed by the Maimonides researchers is appropriate. Moreover, he is right in recognizing that a uniform definition of the dependent variable must be decided upon if the significance of the Maimonides studies collectively is to be determined.

Although Child's own analysis, described above, is sound, it has the disadvantage of not including all the studies in the data base. An alternate approach can be taken by recalculating the delinquent Zs using an error term that assumes "worst-case" dependence of judgings. I decided to undertake such an analysis, which thus included all the formal series. I also decided to use a uniform method of scoring (ranks) rather than the most sensitive method given in the report.

My statistical consultant developed a revised Z formula as follows:

$$Z=(T-N(N+1)/2[\pm.5])/(N^2(N+1)/12)^{.5}$$

where T is the sum of ranks assigned to the target and N is the total number of trials. As the number of trials in these studies varies from 7 to 12, the assumption of normality is unlikely to be grossly violated, although marginal outcomes should be interpreted cautiously.

Separate analyses were performed for subject judging and independent judging. In cases where more than one independent judge was employed, the

means of the judges' ranks were introduced into the equation, a slightly conservative procedure. In the later studies, which employed ratings, the sums of the ratings of the multiple judges were converted to ranks for the analysis. Finally, in a handful of cases the only information available was whether the target was a hit or a miss, i.e., above or below the theoretical median rank. In these cases, all hits were assigned the theoretical median of the possible "hit" ranks and the misses the theoretical median of the possible "miss" ranks. (E.g., in an eight-trial series, the hits would all be assigned a rank of 2.5 and the misses 6.5.) This procedure is also conservative.

The Zs computed by the above methods are presented in Table 1. When these Zs are combined by the Stouffer method over all 11 studies, the cumulative Z for independent judging was 5.41. The corresponding Z for subject judging, cumulated over the eight studies which employed subject judging, was 3.09, $p<.005$. Thus, even when one includes the screening studies, the cumulative results of the formal Maimonides dream experiments are clearly significant statistically. As Child's analysis indicates, the pilot sessions (not included in my analysis) do not detract from this trend.

Given that the collective outcome of the Maimonides experiments cannot be attributed to chance, what can be said about the likelihood of these results being attributable to nonparanormal factors?

Sensory Leakage

The most serious allegation here is Hansel's contention that the experimenter in the Van de Castle study appears to have been present with the agent when the latter opened the target envelope. The following is the paragraph upon which Hansel based this inference. I have underscored those phrases which Hansel himself emphasized in his critique and which led him to the inference.

Table 1

Z STATISTICS OF RANKS CORRECTED (WHEN NECESSARY)

FOR POSSIBLE DEPENDENCE OF JUDGINGS*

	Zs	
	Indep. Js	Subj Js
I. Single Trial per Subject		
A. Screening I	0.62	1.24
B. Screening II	-0.21	1.08
C. Sensory Bombardment	3.25	0.00
II. Multiple Trials per Subject		
A. Erwin I	1.64	1.05
B. Erwin II	3.54	
C. Grayeb	-0.51	0.51
D. Posin	1.08	1.08
E. Van de Castle	2.61	2.86
F. Bessent I	2.53	
G. Bessent II	2.96	
H. Rock Concert	0.44	0.92
TOTAL (Stouffer Z)	5.41	3.09

* Underscoring means that judgings were truly independent and the uncorrected sum-of-ranks Z formula was applied.

Upon arriving in his room, A opened the envelope containing the target picture. He was encouraged to write down his associations, to visualize the picture, to concentrate upon it, and to treat it in any other manner which would make its contents a dynamic part of his conscious processes. Once this was done, there was no way the A could communicate with E or with S without leaving his room and breaching the conditions of the experiment." (Ullman & Krippner, 1970; pp. 99-100).

First of all, nowhere is it stated that E accompanied A to his room. "He was encouraged" could be taken to imply this, but it could also be read as implying that the "encouragement" had been part of the general instructions given to A before the experiment began. "Once this was done" could be taken to mean, as Hansel believes, that only after the target had been opened (in E's presence) was A to E communication impossible, but, if the more generous interpretation of the preceding phrase is correct, it could mean that as soon as A entered the room such communication was impossible.

There is no question that the paragraph is ambiguous and poorly worded. However, by no stretch of the imagination is the implication that E accompanied A to his room clear enough to justify Hansel all but concluding that this is what happened. Further, certain aspects of the procedure seem to argue against Hansel's interpretation. Doesn't it seem odd, for example, that E would need to remind A before each trial how to do the sending? Fortunately, the procedure is stated more clearly in one of the other reports of the experiment, where it is affirmed that the experimenter only stayed with the agent until the latter went to his room to open the target envelope (Ullman & Krippner, 1968).

The other possibility alluded to by Hansel concerns cheating on the part of one or more of the participants. The unsuccessful Foulkes experiment with Van de Castle was indeed somewhat more secure in this regard than the Maimonides experiments. In particular, the latter, unlike the former, did not preclude the possibility that the agent might leave his or her room, sneak down the hall, and somehow convey information about the

target to the percipient without the experimenter knowing it. Indeed, the reports of the first screening and first Erwin experiments refer to the agent occasionally relieving the experimenter during the night, although never talking to the subject. However, there is no evidence that an agent ever compromised a session and several agents would have to be implicated if all the significant Maimonides experiments are to be accounted for as fraud. Since the agents in all the Maimonides experiments were lab staff, this specific criticism falls into the category of experimenter fraud, which can be offered as a possible alternative explanation of all the experiments considered in this review.

However, the fact remains that two experiments with different outcomes (i.e., both the Maimonides and the Wyoming experiments with Van de Castle) did differ procedurally in terms of the opportunities they provided for fraud by the agent. However, they differed in other respects as well. Van de Castle (1977) notes, for example, that he was disturbed by the skepticism of the Wyoming team and that this created a bad psychological climate for the Wyoming experiment. The Wyoming investigators indeed reported evidence of negative feelings toward the experimenters in Van de Castle's dreams during the experiment. Critics often complain bitterly that parapsychologists use this kind of argument as an alibi to explain away failures after the fact. It certainly would be premature to conclude that Van de Castle's explanation is the correct one, but the fact remains that the psychological state of the subject differed in the two experiments and that this was as real a difference as the procedural differences stressed by Hansel. Also, <u>if</u> a "psi" process does exist, it is not unreasonable to suppose that it is influenced by the psychological state of the percipient. Other differences, such as a higher concentration of sessions in the Wyoming experiments, could also have been factors. In short, as long as multiple differences in conditions exist, one cannot confidently attribute differences in outcome to any one of them, especially since none is favored

by other internal evidence in the data.

Lack of Baseline Judgings

The artifact which baseline judgings are supposed to control for, as is clear from reading Alcock's (1981) critique, is the possibility that a target-dream correspondence will be considered evidential because of chance correspondences or because the dream protocols contain highly general statements which could apply to many pictures. The Maimonides judging and analysis procedures in fact control for this artifact because the rank or rating the dream protocol receives depends on how closely it corresponds to the target picture relative to how well it corresponds to the other pictures in the judging pool or set. To put this another way, the mean ratings or rankings assigned to the incorrect pairings serve as the baseline against which ratings and rankings assigned to the correct pairings are assessed.

Another way to address this issue is to ask what the interpretation would be if control judgings in which the correct pairings were assigned randomly or arbitrarily consistently yielded significant results. Such an outcome would be every bit as anomalous as that of the real Maimonides experiments and would fit many definitions of psi, including the one used for this review. If the outcome, on the other hand, were nonsignificant, its deviation from the theoretical "chance" value is properly construed as error and thus should not be incorporated into the baseline estimate. In other words, for this type of research problem, the best external baseline is the theoretical estimate built into the Maimonides procedure.

Many psi experiments other than the Maimonides dream experiments compare obtained results to theoretically defined baselines. The same basic arguments apply in those cases. For a further discussion, see Palmer (1982).

Randomization

A potential source of bias not addressed by previous reviewers of the Maimonides experiments is the inadequacy of the randomization procedure used to select targets in some of the experiments. For example, in the second Erwin experiment, a random digit was used to select which of the ten art prints in the pool would be the target for the first trial. The same procedure was used for subsequent trials, except that if the random digit exceeded the number of prints remaining in the pool, the selector would go back to the first print and continue counting until the random number was reached. A moment's reflection will reveal that this procedure does not lead to each print having an equal opportunity of being selected for each trial. For example, for the second trial, selection of a random digit "1" or "0" ("0" being equivalent to "10") leads to the first print being selected, whereas each of the remaining prints are associated with only one digit; i.e., the first print has twice as much chance of being selected for this trial as any of the others. A proper procedure would have been to select a new random digit each time a digit exceeded the number of prints in the pool.

To determine the extent of the bias, I performed a computer simulation of the above selection procedure. The random numbers were determined by a random event generator, and 1000 mock "experiments" were run, each consisting of eight trials with an initial pool of ten prints as in the second Erwin experiment.

The resulting matrix is reproduced as Table 2. The figures inside the table refer to the number of times each print was selected for each trial. Eight chi-squares were also computed, one for each trial, to indicate the extent to which the distribution of selections for that trial departed from the ideal of each print being selected an equal number of times.

The chi-square for the first trial was not significant. This is to be expected, because the procedure is adequate for the first trial. However,

Table 2

RESULTS OF COMPUTER SIMULATION

TESTING FOR BIASED TARGET SELECTION

		TRIAL							
		1	2	3	4	5	6	7	8
TARGET	A	82	<u>205</u>	142	101	91	78	96	79
	B	97	90	<u>150</u>	144	113	83	97	96
	C	91	101	<u>145</u>	121	119	86	87	109
	D	102	84	86	<u>135</u>	98	83	111	105
	E	87	84	83	120	<u>126</u>	101	108	98
	F	108	87	95	88	<u>111</u>	96	108	101
	G	102	92	72	79	<u>111</u>	106	101	109
	H	115	84	75	69	<u>104</u>	103	81	109
	I	116	85	72	71	65	<u>125</u>	112	104
	J	100	88	80	72	62	<u>139</u>	99	90
	$\underline{X}^2$	11.36	124.96	93.92	71.74	42.18	34.66	9.50	8.46

Numbers indicate the number of times in 1000 "experiments" that target was selected for that trial; underscored numbers show the maximum frequency for the target

the chi-squares for Trials 2 through 6 are highly significant ($p<.001$). The bias is strongest in Trial 2 and steadily declines until by Trial 7 it is no longer detectable for the sample size employed.

The most important feature of the bias is that within Trials 2 through 6 there is a tendency for earlier members of the target pool to be favored on the earlier trials and later members on the later trials. This can be seen by observing in Table 2 for which trial (in the range of Trials 2 through 6) each print receives its maximum number of selections. These figures are underscored in the table and form a virtual diagonal from the upper left to lower right. For instance, Print A receives its maximum number of selections on Trial 2, whereas Print J receives its maximum number of selections on Trial 6.

This bias is serious to the extent that the judge has a tendency to assign early targets in the pool to early trials, either as a natural tendency or because of knowledge that such a bias exists in the randomization procedure. Fortunately, in the second Erwin experiment the judges were all asked to evaluate the possible target-transcript pairings in random order. If this means that they had no knowledge of the original ordering of the targets (i.e., the order of the envelopes before the first trial), then the bias can be considered irrelevant, unless one entertains the rather implausible assumption that the order of the subject's dreams was somehow naturally correlated with the order of the targets in the pool. Even if the judges did know the target order, the fact that they judged the pairs in random order might tend to neutralize any <u>natural</u> judging biases toward selecting one of the first targets seen for early trials, and so on. Randomization of targets given to the judges was not discussed in the reports of the first Erwin experiment. However, Krippner (personal communication) claims that in all the experiments targets were given to each judge in a different random order. How this randomization was accomplished is not stated.

The biased selection procedure poses a slightly different problem in the Van de Castle study than in the Erwin studies, because a separate target pool was used for each trial. Since each pool contained eight prints, this meant that for each trial the first two members of the pool had twice as great a chance of being selected as the others. Precise details of how the pools were constructed were not given in the report, but if there had been any tendency to put the "best" art prints early in the pool--an unverified but not implausible assumption--an effective bias could have resulted.

It is not clear what target selection procedure was used in the "sensory bombardment" experiment. If the faulty method was used, the bias would be comparable to that which applies to the Van de Castle experiment, since there again a single target pool was used for each trial. It also is not reported what randomization procedure was used in the replications of the Van de Castle and "sensory bombardment" experiments conducted by the Wyoming team. Finally, it should be noted that the faulty target selection procedure was <u>not</u> used in the two successful precognition experiments with Bessent. The procedures used in the second of these experiments, although complicated, seem adequate.

Another form of biased target selection occurred in the first of the precognition studies with Bessent (Krippner et al., 1971), however. In this experiment, a word was randomly selected from a dictionary of common dream themes and one of the experimenters created a multi-sensory experience (like a mini-drama) which Bessent experienced the morning after the test night. It thus served as the precognitive target. Descriptions of these experiences were given to the judges for matching with the dream transcripts.

The problem with this procedure is that even though the <u>topic</u> was selected randomly, the actual material in the description was not. For example, the experimenter could conceivably have been influenced in his preparation of the experiences by information he had innocently acquired

about Bessent's activities or thoughts during the previous day, or by newsworthy happenings that day totally unrelated to Bessent. If such activities or events had come to be reflected in Bessent's dreams, artifactual correspondences could have been produced.

None of the biases discussed in this section seem particularly likely as explanations even of the experiments to which they apply, because they require the acceptance of rather implausible ad hoc assumptions. Nonetheless, they must be treated as possible explanations of the results.

NOTE

[1] Unless noted otherwise, p-values cited in this report are two-tailed. In general, I have cited the p-value given by the authors when referring to tests they computed. I have generally cited two-tailed probabilities for my own analyses. Z-scores which exceed 4.0 are generally considered sufficiently astronomical to not require the citation of the exact p-value alongside them.

Chapter 3

REMOTE VIEWING

A great deal of public attention has accrued to experiments using a free-response ESP procedure called "remote viewing (RV)." The main distinguishing characteristic of this procedure is that the targets tend to be "real" objects or geographical sites as opposed to photographs, slides, etc. However, it is also likely that the term was adopted to avoid the "occult" connotations which, despite the efforts and wishes of conservatives like Rhine, have become attached to the term "ESP."

The remote viewing procedure is most closely identified with two physicists, Harold Puthoff and Russell Targ, who at the time of their initial RV experiments were both employed at SRI International in California. This background and affiliation is part of the reason that their research has attained such notoriety in scientific circles.

I will begin by critically reviewing the primary RV experiments of Puthoff and Targ and the controversy about these experiments initiated by psychologists David Marks and Richard Kammann. I will then critically discuss the major replication attempts by Bisaha and Dunne, Schlitz and Gruber, Karnes, and Marks and Kammann. I will not consider various minor experiments, especially those using the "group remote viewing" procedure in which multiple subjects attempt to reproduce a single target.

Puthoff and Targ Experiments

The experiments to be considered used a total of nine subjects, three of whom were labeled as "experienced" (i.e., having participated and succeeded in previous psi experiments), three as "learners" and three as "visitors." The most extensive testing and the most successful (and controversial) results were associated with a former police commissioner named Pat Price and a professional photographer named Hella Hammid

(Puthoff & Targ, 1979).

Main Series

A pool of over 100 target locations within a short driving distance from SRI was assembled by a person not otherwise involved with the experiments. This person randomly selected one location from this pool to be the target for each trial. The method of randomization was not specified nor is it clear whether targets were placed back in the pool after they had been used (i.e., sampling with replacement).

For each trial, a group of two to four "outbound experimenters" ascertained the target location and drove to it. They then observed the location for 15 minutes, during which time the subject (who was located at SRI with the "inbound experimenter") attempted to receive impressions of the site. These impressions were recorded on tape and the subject also drew sketches of the presumed target. The inbound experimenter, who was himself blind to the target location as well as to the contents of the pool, asked the subject questions in an effort to achieve further clarification and elaboration of the impressions. Following the trial, the subject was taken to the site for feedback.

The total of 39 trials was divided into five groups of five to nine trials, each group consisting of the attempts of one or two subjects. For each trial, an unedited transcript of the subject's tape-recorded impressions was attached to the subject's sketches. (Hereafter, the term "transcript" will be defined as including these sketches.) The transcripts for each group of trials were assembled and given to one outside judge who was asked to visit each of the target sites for that group and rank the transcripts in the order of the degree of perceived correspondence to the site. The ranks assigned to the correct transcripts for all trials in the group were then summed and the sum was evaluated for statistical significance by reference to exact probability tables developed by

Morris (1972).

Results. Four of the five groups of trials yielded one-tailed probabilities of less than .05 in the psi-hitting direction. The most significant groups were the nine trials of Price ($p = 2.9 \times 10^{-5}$) and the nine trials of Hammid ($p = 1.8 \times 10^{-6}$).

Technology Series

The target pool for this series, which was designed to assess the resolution capacity of RV, consisted of seven pieces of equipment: drill press, photocopy machine, video terminal, chart recorder, random event generator, machine shop, and typewriter. It was specified that sampling from the pool occurred with replacement. Otherwise, the randomization procedure was the same as in the geographical series. The test and judging procedures were also the same as those previously employed, except that only the subjects' sketches were used for judging.

Twelve trials were completed by five subjects, all but one of whom had participated in the geographical series. Multiple responses to a given target were combined for judging, thereby reducing the number of trials for judging from twelve to seven. The sum of ranks given to the correct targets was again evaluated for significance by Morris' tables.

Results. The total sum of ranks was 18 ($p<.05$, one-tailed) in the psi hitting direction.

The Marks-Kammann Critique

Sensory Cues. In their book Psychology of the Psychic, Marks and Kammann (1980) leveled a harsh critique at the Puthoff-Targ RV experiments. Their most important argument concerned the availability to judges of sensory cues from the unedited transcripts of the subjects' impressions. Marks and Kammann were able to gain access to the raw records of the Price and Hammid series. In each case they noticed that the transcripts contained

information that could help the judge match them correctly to the target list, provided (in most cases) that the target list was not randomized, thus allowing the judge knowledge of the correct target order. For example, the fourth transcript in the Price series contained the statement from the inbound experimenter: "Nothing like having three successes behind you." This statement could cue the judges that the trial was the fourth one in the series, or that it certainly did not occur earlier than that.

Marks and Kammann then cited a letter written by the judge (who was their source for both the raw data and the letter) to the effect that for both the Price and Hammid series he had received the list of target locations in the order that they had been used, i.e., unrandomized. The lack of target randomization for the Price series was acknowledged by Puthoff and Targ (1981) but was challenged both by them and by Morris (1980) with respect to the Hammid series. Morris, who had requested and received a copy of the judge's letter, noted that the judge explicitly stated that he did not know whether the target list had been randomized or not and thus decided to (re)randomize it himself. While observing that the experimenters should have told the judge explicitly that the list had been randomized, Morris concluded that the judge's letter refuted the assertion that Marks and Kamman had made about the Hammid series.

In rebuttal, Marks (1981a) did not directly challenge Morris' assertion. However, he did provide additional evidence in support of his basic argument. The judge's letter revealed that in addition to the target list itself, he had received two other sources of information about the targets for the Hammid series. One of these was pages of notes each containing information about the target site for that trial. He had discovered after judging that the order of these pages correlated .83 ($p<.01$) with the order of target usage. This "almost perfect" (p. 199) correspondence, Marks argued, could have provided sensory cues to the judge. The other was a map of the area designating the target sites. In a

subsequent paper, Marks (1982) reported that a judge of his choosing was able to use information provided by the map plus cues in the transcripts to obtain a more significant score than the original judge on six transcripts from the Hammid series. (For some reason, Marks was not permitted access to the other three.)

With respect to the Price series (for which the availability of sensory cues was conceded), Marks and Kammann (1980) sought to demonstrate empirically that the cues could account for the significant results of this series. First, eight judges were given a list of the targets in the correct order, as well as a randomized set of transcripts containing only the biasing statements from the corresponding real transcripts. On the basis of this information alone, and without actually visiting the target sites, each of the eight judges was able to match the targets and transcripts to a highly significant degree. Thus, the cues indeed had the potential to bias the judging.

However, the crucial point is whether the matching could be performed successfully with the biasing cues removed. To determine this, two additional judges, described only as "research psychologists" (p. 30), were asked to rank the list of targets in random order against randomized transcripts identical to the originals except that the biasing cues had been removed. These judges actually visited the sites. Since four of the trials had been published and the judges might have seen the pertinent information, this analysis was restricted to the remaining five trials. The matchings of each judge were nonsignificant and close to chance expectation. The authors thus concluded that "...the successful identification of target sites by judges is impossible unless multiple extraneous cues...available in the original unedited transcripts are utilized" (Marks & Kammann, 1978).

Charles Tart (Tart, Puthoff, & Targ, 1980) attempted to counter this criticism by editing all nine transcripts, "removing all phrases suggested as potential cues by Marks and Kammann, and...any additional phrases for

which even the most remote post hoc cue argument could be made," and having them matched against the nine randomized target locations by "a new independent qualified judge (having previously shown competence in...analysis of similar materials) who was unfamiliar with the Price series." This judge achieved a high level of significance comparable to that obtained in the original analysis.

Marks (1981b) objected that the transcripts had been edited by Tart, who knew the correct matches and who could have been biased in his editing. Second, he questioned whether it could be established that the "blind" judge really lacked access to information about the four published trials. Puthoff and Targ (1981) then countered by briefly describing a second reanalysis for which the probability of a hit for each trial was adjusted to account for the biasing effect of cues and the revised p-value remained highly significant. I was unable to find details of this analysis either in this report or in a supposedly more detailed paper referred to therein.

Marks and Kammann were unable to gain access to raw data from the other RV series; thus, they had to resort to speculations about how possible breaches of protocol (e.g., informal contacts between the experimenter and the judges) could have biased the series even if cues had been removed and/or the data sheets properly randomized.

Data Selection. The second major criticism in Psychology of the Psychic referred to data selection. Although Puthoff and Targ claimed in their popular book Mind Reach (Targ & Puthoff, 1977) that they had not selected only their best results for publication, Marks and Kammann claimed to find circumstantial evidence to the contrary. They noted, for example, that in Mind Reach Targ and Puthoff referred to "more than one hundred experiments of [remote viewing]" (pp. 9-10) whereas only 55 had been published. They suggested that unsuccessful experiments were labeled as "demonstrations" after the fact and were dropped from consideration.

Concrete examples of indirect evidence that some trials were omitted were provided in the case of the Hammid and technology series. These will be discussed below.

Evaluation of the Marks-Kammann Critique

Sensory Cues. The biasing information uncovered by Marks and Kammann in the unedited transcripts of the Price and Hammid series clearly render the results of the original judging of these series invalid. Fortunately, the experiments were designed in such a way that a proper rejudging could easily be conducted. No such rejudging has been attempted for the Hammid series. Two attempts were made for the Price series, one yielding significant evidence of RV and the other yielding chance results. Unfortunately, neither rejudging completely excluded the possibility of bias.

Two major problems beset the Marks and Kamman rejudging. First, by eliminating the four published trials, they drastically reduced the power of their statistical test, thereby making it more difficult to reject the null hypothesis. Moreover, since the best matches tended to be the ones selected for publication, those trials retained for analysis were not truly representative of the whole data set. This problem is illustrated by the results of Tart et al.'s rejudging; the p-value they obtained based on judging all nine transcripts was 10^{-4}, whereas that based on just the five transcripts selected by Marks and Kammann was only .025. Surely it would have been possible for Marks and Kammann to find a judge or set of judges for whom familiarity with the RV experiments could have been reasonably excluded; in fact, the judges would not even have needed to be informed that the transcripts pertained to an ESP experiment at all, and they could have been asked afterwards if the material looked familiar.

A potentially more serious problem, however, involves the selection of judges by Marks and Kammann. An obvious and important qualification for judges in this type of experiment is that they be highly motivated to

achieve correct matches. Otherwise, they may give the task short shrift or be less observant. When the judges are selected by persons favorably disposed to the experimental hypothesis, it is reasonable to assume that this qualification is met. That is not the case when the person selecting the judges is a skeptic. Under such circumstances the reader requires additional assurances about the judge's motivation. Suspicion is particularly justified in the present case because the judges were "research psychologists," a population that is notoriously hostile to parapsychology. The burden is on Marks and Kammann to provide the necessary assurances on this point.

The two problems with the Tart rejudging noted by Marks and Kammann are not as serious as those affecting their own rejudging, but they are troublesome nonetheless. It is indeed possible that Tart, who knew the correct matchings and was motivated to see the RV hypothesis confirmed, might unwittingly have been biased to more readily excise statements from the transcripts unrelated to the target than statements related to the target, especially since a liberal exclusion rule was employed. (It is not reported who edited the transcripts in the Marks and Kammann rejudging, so this problem might apply in their case as well.) Second, further assurances about the blindness of Tart's judge would be desirable.

Finally, some comments are in order about the extent of potential bias in the original judging of the Hammid series. Here it seems that the main target list was randomized, but questions were raised about the accompanying information pages and the map of the area.

The close correspondence between the ordering of the information pages and the order of target usage, although problematic, is not quite as damaging as Marks and Kammann imply. While a .83 correlation seems high, it in fact represents only about 70% of the variance. For example, I had no difficulty generating a sequence of pages correlating .85 with the sequence of target usage such that none of the nine pages was in their "correct"

locations. However, if the judge had been sensitive to the possibility of the pages being in the correct order, he could have inferred rather reliably whether a target had been visited during the first or second half of the experiment, which would have been useful information. Finally, it should be stressed that the .83 correlation does not suggest that the experimenters failed to randomize the pages, only that the randomization was done poorly.

The more damaging criticism derives from the success of Marks' judge in performing a significant matching of transcripts and targets with the aid of cues from the map. A curious feature of this analysis is that the judge's success seemed attributable in part to the apparent validity of an assumption he made that targets close together on the map were visited by the outbound experimenter on successive trials on the same day. But if the targets for each trial were selected randomly, as stated in the published protocols, the locations of successive trials should have been independent of their physical proximity to each other. Does this mean that the judge was "lucky" enough to a gear his judging to a freak correspondence, or does it mean that the published protocol was not really followed?

In conclusion, although the RV researchers have succeeded somewhat in neutralizing the Marks and Kammann critique pertaining to sensory cues, legitimate grounds for doubt remain about the evidentiality of the data. Fortunately, the validity of the sensory-cue criticism could still be resolved by means of a further rejudging of all the series in the experiment which had the following characteristics: (1) editing of the transcripts by an impartial person blind to the correct matchings; (2) adequate randomization of all judging materials; (3) inclusion of all trials; and (4) judging of the edited transcripts by one (preferably more) judge(s) who are (a) highly motivated to achieve correct matches, (b) demonstratably unlikely to have information about the RV experiments, and (c) uninformed about the identity of the data they are to evaluate.

Data Selection. Marks and Kammann strongly suggest that Puthoff and Targ selectively published only positive results, thereby misrepresenting their actual rate of success. Consider two examples from the Hammid series discussed in Psychology of the Psychic (pp. 34-35). Although Marks (1982) later retracted their charge as it pertains to this series (based upon further information received from Puthoff), an analysis of it may nonetheless be useful in revealing the kinds of ambiguities and gratuitous interpretations of these ambiguities that have beset the remote-viewing controversy from its inception.

First, Marks and Kammann cited a statement made by the inbound experimenter from the transcript of Trial 4: (Targ): "Hella [Hammid] has made a drawing of Hal's [Puthoff] first location. And we'll see where he is for the next fifteen minutes." According to Marks and Kammann, "This [Hal's first location] is clearly [italics mine] a reference to the preceding experiment...in which Hal Puthoff had visited the target site." But since Targ had been the outbound experimenter for Trial 3, Marks and Kammann concluded that there must have been an unreported trial between 3 and 4 for which Puthoff was the outbound experimenter.

In my judgment, it is anything but clear, at least based on the Marks and Kammann account, that the quotation refers to any preceding trial. It seems much more plausible that the statement refers to the current trial (for which Puthoff must have been the outbound experimenter, since Targ was the inbound experimenter) and that Hammid had made her sketch for that trial before recording her verbal impressions. If there is something else in the transcript that made it "clear" to Marks and Kammann that this was not the case, they have done a disservice to their position by not stating it.

In the next paragraph, Marks and Kammann quote the following statement from the last trial in the series: "Hal has gone off to the first of three remote sites that he will visit in the experiment." On the basis of this statement they imply that there were at least two unreported trials in the

Hammid series. The problem here is that Puthoff and Targ consistently use the term "experiment"--not in the sense of a series-- but as a synonym for what most parapsychologists call a trial. Thus, Targ was most likely saying that Puthoff visited three sites in the same trial, not that there were three trials. Marks and Kammann should be aware of this usage, because in their book they frequently quote statements by Puthoff and Targ which adopt it.

Even if we assume that Targ was using the term "experiment" as a synonym for series, the Marks and Kammann interpretation makes no sense. Taken that way, the statement says that the Hammid series consisted of three trials, when according to the Marks and Kammann rendition it consisted of at least nine and probably 13 trials.

None of this is meant to take away from the fact that the statement is puzzling and ambiguous. Why, indeed, should the outbound experimenter visit three locations on the same trial? Could it refer to the fact that the outbound experimenter positioned himself at three different locations at the same (broadly defined) site? The statement could eventually prove troublesome and Puthoff and Targ owe us an explanation. The point, however, is that Marks and Kammann had no grounds for jumping to the conclusion that the statement is evidence of data selection.

A final example of Marks and Kammann's jumping to unwarranted conclusions occurs in their discussion of a trial from the technology series. They imply that data selection was the reason that in a secondary analysis (not the primary analysis described in a previous section) a judge was shown only the better of the two responses to the drill press target. They fail to appreciate that the objective of the analysis was not to evaluate the significance of the trial per se but to demonstrate that the better response was so accurate that the judge could not only match the target but, based upon the drawing, correctly name it. This intent is admittedly not explicitly stated in the report but a careful reading causes

it to emerge as a likely possibility.

The above discussion is not meant to imply that there are not questions pertinent to the possibility of data selection that need to be answered. For example, it would be desirable to have a complete list of the "more than one hundred [RV] experiments" (p. 9) to which Targ and Puthoff refer in Mind Reach. On the other hand, the conclusion reached by Marks and Kammann in Psychology of the Psychic that "there is clear evidence that [data] selection has occurred" (p. 41) is unwarranted and, especially given the severity of the charge (which amounts to an accusation of experimenter fraud), unfair.

Other Criticisms

Logical Inference. 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. However, more precise information on this point would have been desirable. This criticism does not apply to the technology series, where sampling occurred with replacement.

Statistics. The Morris tables used by Puthoff and Targ assume statistical independence of trials. The important point is not the independence of the actual trials as they occur but instead whether the judge treats the trials as independent during judging. For example, the assumption of independence would be violated if the judge were reluctant to give a ranking of "1" to a transcript which he had already ranked as "1" against another target. It is plausible to assume that a judge might be

tempted to do this, especially if he knows that the target pool was sampled without replacement. In any event, there are no indications that the judge was admonished to make his ratings independently.

As a result, it is likely that the original published significance levels are biased. However, the point is moot because the data were later reanalyzed using a direct-count-of-permutation method suggested by Scott (1972) in which the p-value corresponds to the number of possible permutations of the matrix of ranks that would yield a lower sum of ranks than that actually obtained. This method takes into account the possible nonindependence of rankings. The p-values obtained by this method closely approximated those obtained by the earlier method, with five of the six series continuing to be significant by a one-tailed test (Puthoff, Targ, & May, 1979).

Attempted Replications

In this section I will critically review the research of the four major replicators of the Puthoff and Targ remote viewing studies. John Bisaha and Marilyn Schlitz have consistently obtained positive results; Edward Karnes and Marks and Kammann have consistently obtained negative results. Although each has undertaken multiple experimental series, I will focus primarily on the most prominent single experiment of each investigator. With the possible exception of Karnes, the methodology has been fairly uniform within experimenter. In discussing methodology, I will focus on those aspects in which the procedures differed from those adopted by Puthoff and Targ.

Bisaha and Dunne

In collaboration with Brenda Dunne, Bisaha obtained statistically significant evidence of RV in three experiments (Bisaha & Dunne, 1979; Dunne & Bisaha, 1979). The most prominent of these experiments (Dunne & Bisaha, 1979) used a precognition procedure in which the subject was asked

to describe a location to be selected by the outbound experimenter five minutes after the response period was terminated. The target pool consisted of 100 sealed envelopes, each containing the name of a location in the Chicago area. For each trial, a sub-pool of 10 of these envelopes was chosen by an unspecified random method. The actual target was selected by the outbound experimenter from among these ten by reaching into a container and picking out one of ten equally sized folded sheets of paper. The appropriate envelope was then opened and the target location revealed. The outbound experimenter was given 15 minutes to get to the target site, where she remained for 15 minutes, taking a photograph of the site as well as making notes about it. The subject received an unspecified form of feedback about the identity of the target site after each trial.

Two inexperienced subjects completed a total of eight trials. A single judge was assigned to each of the eight target locations. The judge was given a photograph or photographs of the target site along with its name and the outbounder's notes made at that site, and then was asked to rank the eight unedited transcripts in order of their perceived similarity to the target. Judges did not actually visit the target sites. The sum of ranks assigned to the correct transcripts was evaluated by an expanded version of Morris' tables and found to be significant ($p<.008$, one-tailed).

Criticisms. Marks (1982) made three critical points about the Bisaha experiments. The first point was that results from only seven of ten trials were reported. The implication seems to be that the results of the omitted trials were dropped because they were poor; in other words, data selection. The second criticism concerned the editing of the transcripts. Marks obtained the transcripts from Bisaha and found that they did obtain some biasing cues, such as the name of the percipient and the date. The third criticism was that not all the photographs taken of a given site were presented to the judges; if the person who made these selections was not

blind to the transcripts, he might have selected the photograph(s) which gave the best match, thereby biasing the results.

Evaluation. The use of a separate judge for each trial was an important improvement over the method employed by Puthoff and Targ. Not only did it assure independence of the trials in judging, thus allowing proper use of the tables, but it precluded the kinds of intra-judge comparisons across trials that provided the raw material for the sensory-cue criticism of Marks and Kammann (that is, if one can assume that the judges had no way of inferring to which trial in the sequence they had been assigned). However, their choice of the outbounder's photographs as target material for the judges, as opposed to having the judges visit the sites themselves, provided a compensating opportunity for sensory cues. As noted by Stokes (1978), factors such as the weather could be indicated both in the subject's transcripts and either in the notes of the outbound experimenter or the photograph of the site, thereby providing the judge with biasing information. Dunne and Bisaha noted this objection in their report (Dunne & Bisaha, 1979) and said that they had examined the transcripts and photographs for cues and had found none. (No mention was made of the notes.) An independent evaluation excluding the notes would be desirable, however.

The authors refuted the suggestion of logical inference based upon feedback (see p. 52 above) by noting that their target pool was not sampled in a "closed-deck" manner. However, if I understand the sampling procedure correctly, it was not possible for a target to be selected for more than one trial, thus rendering the interpretation possible in principle. However, the size of the overall pool and the fact that no effort was made to force diversification of the sites suggests that this was unlikely to have been a serious source of bias.

Another weakness in the authors' report is a failure to document fully the randomization procedures used in selecting the targets and in preparing

the judging materials. The fact that the judging materials were randomized at all is mentioned only in a preliminary report of the experiment (Bisaha & Dunne, 1977).

The Marks (1982) critique raises some novel and important points, but again it suffers from gratuitous interpretations of the Dunne and Bisaha report. One complicating factor is that the protocol he describes is not that of the main experiment he cites as his reference and the one I have chosen for review (Dunne & Bisaha, 1979), but that of an earlier study (Bisaha & Dunne, 1979). For example, the number of trials in the Dunne and Bisaha experiment is eight, whereas Marks cites it as seven, the number in the earlier experiment. However, the basic arguments apply to both experiments and so the discrepency does not present a serious problem. I will base my evaluation on the later experiment, however.

Nowhere do the authors state or imply in their report of the experiment that there were unreported trials. It is true that each target pool contained ten targets, but this does not necessarily mean that ten trials were planned. For instance, target pools where the number of targets exceeded the number of trials were also employed in the Maimonides dream experiments, where it was clearly stated that the number of preplanned trials was less than ten. However, the authors can be criticized for not reporting whether eight was the number of <u>preplanned</u> trials; thus, optional stopping is a logical possibility.

Mark's final criticism, concerning editing of the transcripts, resembles that of Stokes discussed above. The only examples of biasing information which Marks reported concerned the name of the subject and the date. This kind of information would only be helpful, however, if corresponding information appeared on the target photographs or in the notes. Dunne and Bisaha stated that the photographs were examined for such cues, but no mention was made of this being done for the notes. However, leaving such cues on any of the target material would have be to classified

as gross incompetence for which there is no independent evidence.

It is not clear from the report whether Marks' assertion that not all the photos of each site were given to the judges is valid. The procedure is described by Dunne and Bisaha (1979) as follows:

> Each judge was given one transcript of a percipient's description to read and was then presented with a set of eight photographs with accompanying agent's notes, one of which was the correct target. The number of photographs for each target varied, depending on the agent's judgment of the complexity and size of the target as well as her own observational perspective at the time of the trial. The judges were given these photographs taped to a sheet of paper with the name of the target and the agent's descriptive notes typed below the photographs. (pp.20-21; my emphasis)

The first underlined phrase supports Marks' interpretation. The second underlined phrase could be taken as referring to the sentence immediately preceding it, in which case it would be more consistent with the opposite interpretation, i.e., that all pictures were included. This would require the additional assumption, however, that "eight photographs" in the first sentence really means "eight sets of photographs." If Marks' interpretation is correct, and selection of the photos was made by a person not blind to the transcripts, a bias could have occurred. Although I think his interpretation is the most likely, some doubt remains.

Additional Trials. Dunne, Jahn, and Nelson (1983) subsequently reported the results of 300 separate remote viewing trials, which included the trials just discussed. As procedural details of the subsequent trials are not included in the report, a methodological critique cannot be undertaken. The principal objective of the report was to illustrate the use of a method of analysis for RV data in which both the transcript and the target site are coded on 30 descriptive characteristics (e.g., indoors vs. outdoors). Various scoring schemes based on how well the codings of transcript and site match up on a given trial were assessed statistically with respect to null distributions of scores derived by Monte Carlo methods.

The composite Z-scores derived from this method for the total sample ranged from 5.45 to 7.71, depending upon the particular scoring scheme employed. Results did not seem to be affected by the distance between the viewer and outbound experimenter, nor the time interval between sending and receiving. Twenty-one percipients and nine agents contributed to the formal data base of 168 trials. The fact that only four percipients (19%) and one agent (11%) contributed net negative Z-scores (three or more negative out of the five scoring schemes) suggests that the effect is distributed fairly uniformly across the sample.

Schlitz

Marilyn Schlitz has reported two successful remote viewing experiments using herself as subject (Schlitz & Gruber, 1980, 1981; Schlitz & Haight, 1984). The most prominent of these two was an experiment conducted with Elmar Gruber in which Schlitz was located in Detroit and Gruber (the outbound experimenter) in Rome. A target pool of 40 geographical sites in Rome was intentionally constructed so as to contain several targets of a given type (e.g., fountains, churches). Gruber selected by means of a random number generator one of these 40 sites (without replacement) as the target for each of the ten trials. Gruber visited each site at the time Schlitz was making her response, tape-recording his impressions of the site.

Schlitz, who received no feedback about the target sites until the experiment had been completed, mailed her written impressions and sketches for each trial to Gruber at the end of the experiment. Gruber and another person, who was not aware of the target sequence, translated the subject's transcripts into Italian. They also looked for biasing cues of the type indicated by Marks and Kammann but found none. The translations were then checked for accuracy by a third person who was blind to the target sequence.

Photocopies of the ten transcripts were given in different random orders to five judges who were allowed to visit the sites in any order they

wished. They also had access to Gruber's notes about each site. Not only did they rank the transcripts at each site but they also rated each possible transcript-site pairing for correspondence on a 0-100 scale. The resulting 10x10 matrices of ranks and ratings were each evaluated by the direct-count-of-permutations method and both were found to be significant at approximately 5×10^{-6}, one-tailed.

Concerned about the possibility that the notes might have provided sensory cues of the type referred to in the discussion of the Bisaha experiments, the experimenters later asked two new judges to complete the rankings and ratings without the notes (Schlitz & Gruber, 1981). The results in both cases remained significant, but at a more modest level ($p<.002$).

Evaluation. Although it is not customary for experimenters to serve as their own subjects in psychological research, there does not seem to be concrete objections that can be raised against this procedure in this case. At the same time, some precautions which could easily have been taken either were not taken or not reported. For instance, no mention was made of whether Schlitz sent Gruber the transcripts in random order. Why did Gruber, who knew the target order, allow himself to participate in the translation and editing of the transcripts when two other persons who were blind to the target order were available for the task? Was the order in which the target sites were presented to the judges randomized? Were the judges given the notes in random order?

These problems were eliminated in an otherwise strict replication of the above procedure by Schlitz and Haight (1984). Ten trials were conducted with Schlitz in Durham, North Carolina, and the sender (Haight) in Cocoa Beach, Florida. The response transcripts were edited by a third party who was blind to the correct matchings, both the transcripts and target list were randomized before judging, and there were no notes by the sender. The

results were significant ($p<.05$, one-tailed) for both rankings and ratings. This study is the best controlled and most fully reported of those considered in this chapter.

Karnes

Edward Karnes and associates have reported three nonsignificant RV experiments (Karnes & Susman, 1979; Karnes, Ballou, Susman, & Swaroff, 1979; Karnes, Susman, Klusman, & Turcotte, 1980), although I do not think the Karnes and Susman study should be classified as an RV experiment. Of the remaining two, the 1980 study most closely approximated the SRI procedure and I will focus on it.

The subjects for this experiment were eight self-proclaimed psychics, divided into four sender/receiver pairs according to their own preferences. The four subject pairs completed a total of 16 trials. The target pool consisted of 16 "distinctly different" outdoor and indoor geographical sites, the order of which was randomized by means of a random number table. During each trial, Karnes and the sender went to the target site, the sender taking a movie of the site and recording his or her impressions on tape. Receivers tape-recorded their impressions and drew sketches. The period for both sending and receiving was 15 minutes. Subjects received feedback after each trial.

The sender and receiver mentation reports were transcribed, edited to remove biasing cues, and randomized. Four judges were assigned to each site and asked (1) to indicate the eight transcripts which best described the site and (2) to rank these eight "best" transcripts. In addition to visiting the site, the judge had access to the senders' edited notes and the movie as part of the protocol.

In selecting the eight "best" transcripts, the 64 judges obtained 25 hits (39%) which was lower than the expected 32 hits (50%) to a degree which approached but did not reach significance ($p<.10$). The mean of the ranks

assigned to the correct transcripts was 4.36 compared to MCE of 4.5, which was a nonsignificant difference ($t[24]=.48$).

Evaluation. Although the results of this experiment were nonsignificant, I will nonetheless offer a methodological critique for purposes of comparison with the other studies being considered.

(1)Logical Inference. Since the 16 target sites were selected as distinctly different and subjects received post-trial feedback, the Hyman criticism (regarding the advantage of *avoiding characteristics* of previous targets on subsequent responses) applies. This bias is somewhat ameliorated by the use of multiple subjects, although two of the subjects each completed six trials. Also, randomization methods were not fully documented.

(2)Sensory Cues. These are unlikely, given the randomization of both sender and receiver transcripts and the editing of both. However, it is not indicated whether the person who edited the transcripts was blind to the correct matches. Another possible source of sensory cues, noted by Tart (1980), is that Karnes, who knew the target for each trial, had sensory contact with the receiver during the administration of the instructions prior to the trial. The nonsignificance of results is not an adequate rebuttal to this criticism, since the cues could bias the subject toward incorrect impressions as well as correct ones. On the other hand, it is difficult to see how meaningful cues could be transmitted to the subject during a rather standardized administration of instructions unless one assumes gross negligence by Karnes.

(3)Statistics. The statistical analysis of hits is, strictly speaking, improper, since the judgments of the members of each group of four judges evaluating the same trial were treated as independent. This criticism would

seem to apply as well to the analysis of the ranks. However, it is extremely unlikely that these statistical errors affected the authors' conclusions, since they would more likely tend to inflate rather than reduce significance levels. The analysis of ranks was rather insensitive in that it was restricted to only the best transcripts (i.e., restricted range), but this analysis was secondary.

Marks and Kammann

Marks and Kammann (1980) completed five RV series, totalling 35 trials, in an attempt to replicate the SRI results. One subject participated in each series. The target pool consisted of 100 geographical sites, one of which was selected for each trial, without replacement, by an unspecified random method. The test procedure seemed essentially identical to that used by Puthoff and Targ, and subjects received feedback after each trial. The response transcripts were edited for biasing cues and randomized. There were five judges in each of the first four series and one judge in the fifth. The judges went to each site and ranked the transcripts for that site. There was no statement that the order of sites given to the judges was randomized. The method of statistical analysis was not specified, but the results of each series were reported as nonsignificant.

Evaluation. In criticizing the methodology of the experiment, it is important to keep in mind that Marks and Kammann made a conscious effort to duplicate the SRI method as closely as possible, except for the editing of the transcripts. The one point that should be noted in this connection is that nowhere do the authors state whether the person who edited the transcripts was blind to the correct matchings. In other respects, the same methodological criticisms that applied to the SRI research apply to the Marks and Kammann replication.

General Evaluation

Why have some investigators consistently obtained positive results in RV experiments and others just as consistently obtained negative ones? The traditional skeptical answer to this question is that the experiments which obtain negative results are superior methodologically. However, when we consider the final judgings, the methodological quality of the positive and negative studies reviewed in this report appear to be about equal, at least insofar as quality can be inferred from the experimental reports.

The actual test procedures seem quite uniform, so it is unlikely that the key resides here. Since inexperienced as well as experienced subjects produced positive results in the successful experiments and amateur psychics were used in some of the unsuccessful experiments, subject characteristics also seem to be a poor bet.

The identity of the judges is perhaps a more promising option. Although most of the controversy has focused on the skill of the judges, the motivation of the judges may be a more important variable. Unfortunately, the judges are rarely described in the experimental reports. However, all else being equal, it is reasonable to assume that "pro-psi" experimenters (the ones who achieved the positive results) are more likely to select highly motivated judges than are skeptical experimenters. This is not to suggest that many of the judges used by Karnes and by Marks and Kammann (in their experiment) were skeptics, and the lone judge in their final series was identified as being a "sheep"; yet they might still not have been as highly motivated as the more "successful" judges. Unfortunately, since we lack sufficient data, this interpretation can only be considered an educated guess, at best. The only variable that I can discern which is known to distinguish reliably the results of these experiments is the identity (and the theoretical orientation) of the principle investigators. Why this makes a difference is, of course, the crucial unanswered question.

Some mention should be made of the generally poor level of methodological detail in the experimental reports. This is true of the reports of the skeptics as well as those of the proponents. For example, in only one case (Schlitz & Haight, 1984) was the method of randomizing targets and judging materials described in detail. Often the method was not described at all, and in several key instances it was not mentioned whether any kind of randomization took place. This deficiency of reporting is one of the major reasons why so much controversy has arisen about the RV work and why a proper evaluation of the current status of the evidence is so difficult. However, it would be unfair to single out the remote viewing research in this connection; we will repeatedly confront the problem throughout this report.

Finally, one factor of a more ad hominem nature must be briefly considered in evaluating the SRI research program. I refer to the reluctance of Puthoff and Targ to share their raw data with critical investigators. Marks and Kammann complained bitterly about this in their book and another critic, Christopher Scott, has also had considerable difficulty obtaining the data he needs (e.g., Scott, 1982).

Such recalcitrance inevitably creates the impression that the investigators have something to hide and thus damages the credibility of the research. More importantly, if Marks and Kammann had not been able to obtain the raw data of the Price and Hammid series from the judge, the fatal flaws in the initial judging of the Price series may never have come to light.

On the other hand, one can sympathize with the reluctance of investigators to share data with antagonists who may misrepresent it in print to promote their own viewpoint or ideology or to make public insinuations of fraud based upon inadequate evidence. Regrettably, there is precedence for this type of behavior on the part of critics in the history of parapsychology (e.g., Hansel, 1966, 1980). The nature of some of the

criticisms of Marks and Kammann, as discussed earlier, seems to justify such concerns in the case of these particular commentators.

Nonetheless, given the importance and controversial nature of the SRI research, it is my opinion that the SRI researchers have not been as forthcoming as they could have been in addressing the concerns of critics and, in particular, in seeking independent evaluation of their data and procedures under circumstances that protect their legitimate interests. It is unclear to what extent, if any, external pressures might be responsible for this behavior.

Chapter 4

THE GANZFELD DEBATE

A major preoccupation of parapsychologists since the 1970s has been the exploration of techniques that might be used to increase the manifestation of ESP and PK in experimental settings through the induction of altered states of consciousness (ASCs). One technique that has enjoyed widespread use is the ganzfeld. Originally introduced by Bertini, Lewis, and Witkin (1969), the ganzfeld is a sensory deprivation procedure that encourages inward focusing of attention and a hypnagogic or hypnagogic-like state of consciousness. The principal rationales for its use in parapsychology are that the inward focusing of attention and the reduction of meaningful sensory input facilitate detection of subtle psi-laden mental impressions (Honorton, 1977) and that it reduces "linear constraints on mentation" (Stanford, 1979).

The ganzfeld procedure eliminates patterned stimulation in the visual and auditory modes. Visual isolation is provided by taping acetate hemispheres (halves of ping-pong balls) over the eyes, stuffing small pieces of cotton around the edges, and shining a light through them from a short distance. Many investigators prefer a red to a white visual field, which can be achieved by using either a red light or red-dyed ping-pong balls. Auditory isolation is provided by playing white or pink noise (or a similar stimulus, such as waterfall sounds) to the subject through headphones at moderately loud volume. Specific parameters for the strength of illumination, loudness of sound, etc., have not been standardized and often are not reported. Frequently, subjects are allowed some latitude in adjusting these parameters themselves.

The percipient is left in the ganzfeld for 20 to 45 minutes, although the longer durations are preferred. During this period, percipients are encouraged to observe passively their mental processes. In most cases they

give ongoing mentation reports, although in some studies the reports are postponed until just after the ganzfeld period.

The ganzfeld is used in conjunction with standard free-response ESP test procedures. Targets for ganzfeld sessions consist of relatively complex and often emotionally evocative pictures, usually in the form of photographs or slides. One of these is randomly selected from a large pool to serve as the target for each session (or trial). In most experiments, an agent attempts to send the content of the target picture to the percipient during the ganzfeld period. After the session, either the subject or an outside judge (or judges) assesses the correspondence between the subject's mentation (or a transcript of the mentation report) and each of a set of pictures, one of which is the target, on a double-blind basis.

THE CONTROVERSY

The ganzfeld studies entered into the psi controversy when critic Ray Hyman chose to treat them as a representative sample of state-of-the-art psi research in evaluating experimental parapsychology generally. His choice was conditioned by the fact that strong claims had been made by some parapsychologists (e.g., Honorton, 1978) for the repeatability of the results using this procedure. His interest in this data base evolved into a protracted debate with Charles Honorton (Honorton, 1983; Hyman, 1983). The culmination of this debate was a lengthy exchange which has recently appeared in the Journal of Parapsychology (Honorton, 1985; Hyman, 1985). In the remainder of this chapter, I will first summarize and then critically evaluate this exchange. Two central issues have defined the debate and I will treat them separately. The first issue is the true level of replicability reflected in the data base; the second issue is whether procedural flaws are sufficiently serious to undermine the experiments as evidence for psi.

Issue 1: Replicability

Hyman's Critique

Several published literature reviews have estimated that in the neighborhood of 55% of published ganzfeld experiments have provided significant evidence of ESP (Blackmore, 1980; Honorton, 1978). Hyman chose to base his analysis on an unpublished evaluation by Honorton in which 23 of 42 separate experiments (55%) "achieved overall significance on the primary measure of psi at the .05 level" (Hyman, 1985, p. 5).

In attacking this claim, Hyman set two broad objectives for himself: first, to show that the claimed success rate of .55 was too high and, second, to show that the claimed error rate of .05 was too low.

Success Rate. Regarding the first objective, Hyman made the following points:

(1) Ten experiments in the data base had multiple cells (i.e., experimental conditions), each of which he felt should be treated as a separate experiment. With one exception (to be discussed below), Honorton had pooled cells to arrive at a decision regarding significance for each experiment. Hyman was particularly critical of Honorton's strategy in relation to a very complicated experiment by Braud and Wood (1977) in which the objective was to determine if immediate, trial-by-trial feedback of results could enhance scoring within the ganzfeld paradigm. Among other things, Hyman noted that the baseline condition of this study, which closely approximated a standard ganzfeld experiment, produced nonsignificant results and should be counted as a failed experiment.

The one case where Honorton did not pool cells was an experiment by Raburn (1975) in which the presence of a sender and the percipient's knowledge of whether he was taking a psi test were manipulated in a 2x2 factorial design. Significance was restricted to the one cell in which a sender was present and the subject knew it was a psi test. Honorton chose

not to include the two cells in which subjects were not aware that they were taking a psi test on the grounds that the conditions were grossly atypical of ganzfeld research, and treated the other two cells as separate experiments. Hyman objected to the exclusion as arbitrary, noting that other atypical conditions in other ganzfeld studies were not excluded and that in other ESP studies (but not ganzfeld studies) ESP had been demonstrated when subjects were unaware of being tested for it. He also objected to Honorton's inconsistency in not adhering to his own pooling criteria, which would have resulted in the experiment being classified as a failure even with the exclusion of the two disputed cells.

Hyman concluded that if the cell had been used as the unit of analysis, as he preferred, there would have been 25 "successes" out of 80, thus reducing the percentage of success from 55% to 31%. Details of how this figure was arrived at were not provided.

(2) Hyman claimed circumstantial evidence that a number of mostly unpublished experiments not included in the data base had a lower success rate than those which were included. He first cited studies alluded to in reviews by Blackmore (1980) and Parker and Wiklund (1982) which, if added in, would reduce the success rate from 55% to 43%. Again, no details were provided.

His primary piece of circumstantial evidence was the observation that the rate of successes did not increase as the sample size (and thus the statistical power) increased, as one would expect from statistical theory. He demonstrated his point through an analysis in which he divided the data base into four subgroups of increasing sample size. Estimating the percentage of hits for the entire sample to be 38%, he computed the number of significant experiments expected for the median N of each subgroup and compared it to the observed number of significant experiments in each subsample by a chi-square test. The overall chi-square was highly significant ($\underline{X}^2[4]=31.42$, $\underline{p}<<.001$), the significance being almost entirely

attributable to an excess of significant studies in the smallest class (N<20 trials). As an explanation, he speculated that small studies which were nonsignificant were dismissed as having inadequate power and thus were not reported, whereas larger nonsignificant studies were reported. Also, studies showing poor results on early trials might have been aborted and thus not reported.

Finally, Hyman noted six "retrospective" experiments which he suggested were not originally scheduled for publication but were only published when the results proved significant. One was a compilation of trials conducted when film crews visited the laboratory (Honorton, 1976), one was a classroom demonstration (Child & Levi, 1979), one was not published until three years after it was conducted (Braud, Wood, & Braud, 1975), and the other three were defined as exploratory.

Hyman did not attempt to provide a numerical estimate of how much these factors would reduce the claimed success rate but he concluded this first section of his critique by stating that "the apparent rate of successful replications must be well below 30%" (p. 16).

Error Rate. Regarding the second objective, Hyman focused on the fact that there was no uniform definition of the dependent variable, i.e., the ESP score. In particular, he noted that any of five separate indices (e.g., direct hits, sum of ranks assigned to the targets) were used by the reviewers in classifying experiments as significant. He argued that since the reviewers would accept any of these measures, the ostensible .05 error rate should be adjusted to take account of this fact. To determine the appropriate error rate, he performed two simulations, each consisting of 1000 computer-generated "experiments" with random assignment of scores. Taking into account the observed intercorrelations among the five indices, he concluded that the proper error rate was .22.

He then performed a more conservative simulation in which he determined the error rate for each experiment based upon the number of indices actually reported and came up with an average error rate of approximately .10. However, he provided circumstantial evidence and one concrete example suggesting that not all indices actually used were reported, which, if taken into account, would of course raise the error rate somewhat. He implied that the .22 error rate is the more appropriate, because it is the one that defines the operating assumption of the reviewers. In neither case did Hyman provide enough details about how he arrived at his estimates to allow a proper evaluation.

Hyman concluded by listing five other sources of multiple testing, that each applied to between 7% and 64% of the data base. He felt these should also be taken into account in determining the true error rate. They are: (1) alternative significance tests on the same scores; (2) empirical as well as theoretical baselines; (3) multiple types of ESP scores (other than the indices discussed above); (4) multiple experimental conditions; and (5) use of both subjects and outsiders as judges.

Hyman concluded from his analysis that "the arguments I have made...make a strong case that the overall effective error rate per study could easily be [.25] or higher" (p. 25). Since he had previously concluded that the real success rate "must be well below 30%," his ultimate conclusion that "this rate of 'successful' replication is probably very close to what should be expected by chance given the various options for multiple testing exhibited in this data base" (p. 25) follows naturally.

Honorton's Rebuttal

Success rate. Honorton treated his disagreements with Hyman regarding the appropriate unit of analysis (cell or experiment) and whether two of the four cells of the Raburn experiment should have been disqualified as differences of opinion which cannot be resolved objectively. He invited

readers to "examine the original study descriptions and draw their own conclusions" (Honorton, 1985, p. 54).

Honorton did not dispute Hyman's claim that the success rate of studies with small sample sizes is disproportionately high but he did dispute Hyman's conclusion that this can be attributed to nonpublication of nonsignificant studies with small numbers of subjects. He argued that:

(1) free-response ESP experiments quite often are originally designed for small sample sizes and nonsignificant as well as significant studies of this type are frequently reported;

(2) Hyman's six examples of "retrospective" experiments do not fit his criteria (e.g., two actually had large sample sizes and a third was nonsignificant);

(3) Blackmore's (1980) survey of unpublished ganzfeld studies revealed that none of these studies were aborted solely because of nonsignificant results; and

(4) Hyman has no right to assume that studies with small and large sample sizes are equivalent in other important respects. For example, in two of the larger experiments it was noted that scoring level was inversely related to the number of sessions conducted per day. (However, it should be noted that this is not the same as sample size, although bunching of trials on a single day is more tempting in large studies than in small ones.)

Referring to an analysis to be described below which he claimed corrects for the inflated error rate, Honorton argued that the rate of success of the published experiments is sufficient to compensate for the effects of some unpublished failures. First, he noted that the success rate is not diminished by eliminating the studies with small sample sizes. Second, he cited application of a statistic suggested by Rosenthal (1979) to estimate the number of nonsignificant experiments that it would take to reduce a data base to nonsignificance, given the results of the published experiments. The number of such studies in this case is 423, and it seems

unlikely that publication bias, if it exists at all, is anywhere near that extreme.

Error rate. Honorton conceded the existence of a multiple analysis problem, but he argued that for the purposes of the meta-analysis, it should be restricted to multiple indices or analyses of the overall effect. Analyses which refer to comparisons between conditions or subsamples of the overall study are irrelevant. He then adjusted the significance levels of all 42 studies, based upon the number of indices each employed to test for overall significance, using the Bonferroni inequality (Rosenthal & Rubin, 1984). This reduced the number of significant studies from 23 to 19, or from 55% to 45%.

However, his main rebuttal was to perform a new analysis using as a single, uniform index the number of direct hits. This information was provided for 28 of the 42 experiments. Specifically, he converted the exact probability values of the direct-hit rates to Z-scores and cumulated them according to a method developed by Stouffer (Mosteller & Bush, 1954). The resulting Z was a highly significant 6.60. If one assumes the remaining studies in the data base had an average Z of 0 (chance), the Stouffer Z is slightly reduced to 5.67. He also calculated that 43% of these experiments were significant at the .05 level and that 82% were in the positive direction. Finally, he demonstrated that the Stouffer Z was still significant ($Z=3.67$, $p=.0001$, one-tailed) if the results of experiments from the two most successful laboratories (out of the ten reporting ganzfeld experiments) were eliminated.

Evaluation

Success rate. There is disagreement between the protagonists as to whether the experiment (Honorton) or the experimental condition (Hyman) is the more appropriate unit of analysis in evaluating the success rate of the

ganzfeld, although both acknowledge that a case might be made for either alternative. Not surprisingly, each reviewer chose the alternative that provided the outcome most consistent with his favored conclusion.

Only Hyman really addressed the issue. He offered two reasons why the experiment might be considered the appropriate unit and rejected both. First, he noted that using the condition creates interdependence among the units because the same subjects were often used in different conditions, but he rejected this as a problem because the interdependence exists regardless of how the pie is divided. In other words, the interdependencies are shifted from within the units to between the units. I agree with Hyman that interdependence is not an adequate reason for rejecting the condition as a unit, but neither is it an adequate reason for preferring it.

Second, Hyman conceded that his approach reduced statistical power by reducing the mean unit sample size, but he says this is not a problem because there is no relation between sample size and proportion of significant outcomes in the data base. Here I must disagree with Hyman's reasoning in that I fail to see the relevance of his point to the issue at hand. Whatever the relation between sample size and outcome, dividing a unit into subunits reduces the capacity to detect significance in the subunits compared to that capacity for the whole unit. While some of the power might be recovered by the increase in the number of units analyzed, the issue being debated is not the significance of the overall percentage of successful studies with respect to some total number of studies but the percentage itself: 31% (Hyman) or 55% (Honorton). In my judgment, the difference between these percentages is primarily if not wholly attributable to the relative lack of power of Hyman's procedure compared to Honorton's. Thus, Honorton's procedure is preferable.

I am not persuaded that the "file drawer" problem (lack of publication of nonsignificant studies artifactually increasing the success rate) is nonexistent, but neither am I persuaded that it comes anywhere near

providing the 423 such studies needed to reduce the data base to nonsignificance. In particular, I think Hyman has gone way beyond the data in attributing the relatively high success rate of small-sample experiments to selective reporting.

Error rate. The adequacy of Honorton's rebuttal to Hyman's charge of an inflated error rate now depends upon the adequacy of Honorton's analysis of the direct-hit studies using the single index of number of direct hits. There seems to be a priori justification for adopting this measure, both because it was the one most frequently used in the data base and also because it was the method adopted in the first published report (Honorton & Harper, 1974). However, it was also an advantageous choice for Honorton. For unknown reasons, the mean Z-score of these 28 studies (+1.31) was significantly higher than the mean Z-score of the other 14 studies (+0.01) for which direct-hit scores were not available and other scores were used to compute the Zs ($t[34]=2.06$, $p<.05$). Honorton did not mention this fact in his report, but he did compute a revised cumulative Z-score under the assumption that the 14 omitted experiments had an average direct-hit Z-score of 0, which seems a realistic estimate of the true state of affairs. The cumulative Z was only reduced slightly, from 6.60 to 5.67.

By applying the same reasoning outlined in his discussion of the multiple indices problem, Hyman could challenge Honorton's revised analysis on the ground that many of the individual Z-scores should be reduced to allow for the fact that other indices either were used or could have been used. However, this reasoning is invalid, both in the critique of multiple indices and (if it were to be so applied) in the present case. The kinds of corrections Hyman advocated are appropriate, indeed necessary, if one wishes to draw conclusions from single studies (i.e., does Experiment X, by itself, provide significant support for the hypothesis?). However, in the present controversy the issue is whether a group of experiments, considered jointly,

supports the hypothesis. In this circumstance, the corrections are not appropriate because the p-value one wants from each study is how many times as extreme a score would be achieved in X number of replications (the uncorrected value).

Consider, for example, a sample of 1000 such studies, each of which was analyzed in two ways such that one of these ways always yielded a significant outcome at the .05 level and the other always yielded a nonsignificant outcome. If each of the significant p-values were multiplied by two, as suggested by Hyman, all 1000 of the significant outcomes would be reduced to nonsignificance. Not only would this lead to the absurd conclusion that the effect was a statistical artifact, but the number of truly significant outcomes would actually be significantly lower than the 50 predicted by the null hypothesis ($Z=-7.25$)! This is because the number of significant outcomes expected by chance would not be achieved. Although the example is idealized, the conclusion is obviously absurd.

The Z-score method employed by both Honorton and Hyman in the latter stages of their controversy is much more sensitive than the arbitrary classification of experiments as significant or nonsignificant. An important feature of the data which only the Z-score analysis reflects is the fact that 82% of the direct-hit studies ($Z=3.21$, $p<.01$) and 76% of all studies ($Z=3.24$, $p<.01$) yielded Z-scores in the positive direction. This fact is mentioned only briefly by Honorton and not at all by Hyman, yet it is the most powerful single argument for the non-chance character of the data base. Although it may be quite reasonable to expect publication bias to favor significant over nonsignificant outcomes, it is less reasonable to expect such bias merely with respect to the direction of outcomes.

In conclusion, although the true success rate of all the ganzfeld experiments actually conducted by the time of the controversy is almost certainly less than the 55% originally claimed by Honorton, it is clearly higher than the rate to be expected by chance. Thus, there is something

here to explain. The question remains as to how to explain it.

Issue 2: Procedural Flaws

Hyman's Critique

Hyman began his critique by listing six major categories of procedural flaws that applied to 24%-74% of the experiments in the data base. They can be summarized as follows:

(1) Single Targets: Cases in which the target handled by the agent was included in the set of pictures subsequently judged by the percipient, thus allowing for the possible transmission of sensory cues.

(2) Randomization: Cases in which no randomization or an inadequate method of randomization, such as shuffling cards or tossing coins, was employed to select the order of targets, or cases in which the method was not sufficiently described. Procedures specifying the use of REGs or random number tables (RNTs) were considered acceptable.

(3) Feedback: Cases in which single targets were employed and the order of pictures in the judging set was not properly randomized. Also, cases in which inadequate precautions were taken against the percipient communicating with the agent at the time of feedback.

(4) Documentation: Primarily cases in which it was not reported how frequently the agent was a friend of the percipient or whether this variable affected the results.

(5) Security: Cases where inadequate security was taken against threats to the "validity" of the experiment, particularly cases which employed a single experimenter such that the agent and percipient were not both monitored.

(6) Statistics: Cases in which the statistical tests were improperly applied.

Hyman next responded to the argument of some parapsychologists that flaws can be discounted if it cannot be shown empirically that they make a

difference in study outcome. Conceding, for example, that studies using single target sets did not produce significantly better results than studies using duplicate target sets, he argued that a cause-effect relationship may still exist if this flaw interacts in certain ways with some other flaw that is also causally related to study outcome. Finally, he argued that "such flaws are signs that the study was probably not carefully planned or properly carried out" (p. 31).

Hyman then reported an "exploratory data analysis" in order to "suggest hypotheses about what may be going on" (p. 32). Up to seven measures of psi (e.g., significance, effect size) as dependent variables, along with nine flaws (the six listed in this section plus three reflections of multiple analysis discussed in the previous section) as independent variables, were incorporated in both a cluster analysis and a factor analysis. Each analysis yielded three overlapping clusters (or factors), which Hyman labeled as "general security," "statistics," and "controls." Only the controls factor, which included as the most strongly weighted components the flaws of randomization, feedback, and documentation, correlated significantly with the measures of study outcome. These three component flaws, as well as the statistics flaw, also correlated significantly by themselves with the measures of significance of study outcome.

Hyman then reported a second factor analysis, which included as predictors the five experimenters who contributed the most experiments to the data base as well as the scores in the three clusters derived from the earlier analysis, and a few other predictors. Four factors were extracted, of which Hyman found the fourth the most interesting. It included a high loading on the "controls" cluster (which had previously been a significant predictor of study outcome), plus a high positive loading from one of the experimenters who habitually obtains significant results in ganzfeld experiments and a high negative loading from one of the experimenters who habitually obtains nonsignificant results. Hyman interpreted this result as

supporting the conclusion that "experimenters who pay the most attention to such controls also report the smallest effects" (p. 36), thus providing evidence as well that the "experimenter effect" in parapsychology can be attributed to differences in the care different experimenters put into designing and executing their experiments.

As a final coup de grace, Hyman computed two multiple regression equations using the previously identified significant flaws of randomization, feedback, and documentation as predictors of the statistical significance (Z-scores) and effect size (respectively) of studies in the data base. Both multiple correlations were significant, and the equations predicted that for studies eliminating all three of these flaws, the expected Z-score would be zero and the expected hit rate 27% (assuming a chance hit rate of 25%).

Hyman's final conclusion was that "the current data base has too many problems to be seriously put before outsiders as evidence for psi" (p. 42).

Honorton's Rebuttal

Honorton began by quoting Glass et al. (1981), authors of a seminal book on meta-analysis, in defense of the opinion that the influence of study quality on study outcome is an empirical question that should not be determined a priori. He then proceeded to his main line of defense which was (1) to challenge Hyman's assignments of flaws and (2) to show by his own codings and meta-analysis that when flaws are properly assigned and coded there are in fact no significant relationships between presence of the flaws and study outcome. In order to control for the confounding effects of multiple analysis, Honorton restricted these analyses to the 28 experiments where a direct-hit analysis was possible, using significance levels of this analysis (converted to Z-scores) as his dependent variable.

Honorton agreed with Hyman's codings on two of the six procedural flaws: single target and statistics. On the other variables, his principal

complaint was that Hyman failed to follow his own stated criteria in several instances. With respect to the security, documentation, and feedback flaws, he complained about ambiguity in Hyman's definition of the flaw or inconsistency in his application of it to the studies under consideration. He also questioned the seriousness of some of the "flaws". Specifically, he questioned whether card shuffling is really an inadequate method of randomization in studies where the randomization is performed separately for each trial. Regarding security, he quoted an example showing that a procedure was in fact secure despite the fact that it met Hyman's formal "flaw" criterion of having only one experimenter present.

Second, Honorton computed separate correlations between presence of the flaws single target, randomization, feedback, and documentation with his revised codings, showing in each case that the correlations were not significant. (Hyman had claimed significant correlations for these latter three plus statistics). Finally, for single target and randomization, Honorton demonstrated in each case that the direct-hit experiments which Hyman regarded as adequate with respect to the flaw in question were collectively significant by the Stouffer method. For single target, ten adequate experiments produced a highly significant Stouffer Z of 4.35. For randomization, 16 adequate studies yielded a Stouffer Z of 4.14.

Finally, Hyman's multivariate analyses were dismissed by David Saunders (1985), Honorton's statistical consultant. He noted, first, that the sample of 42 studies was simply too small for factor analysis to be meaningfully employed. Second, the factor analysis was compromised by certain dependencies among the variables entered into them; e.g., each experimenter was treated as a separate variable, guaranteeing that the intercorrelations among these variables must be strongly negative. Finally, the significance of Hyman's final multiple correlations predicting study outcome from flaws are meaningless because the predictors were selected post hoc from at least 84 possibilities.

Evaluation

There is no question that Hyman has uncovered flaws in the methodology of a large percentage of the ganzfeld experiments which could conceivably serve as the basis for conventional interpretations of the results of these experiments, singly as well as collectively. What remains for evaluation is the likelihood that these flaws do indeed account for the results.

Honorton and Hyman disagree about the necessity of empirical support for the conventional interpretation. However, Hyman went to considerable pains to provide such empirical evidence, despite his denial of its necessity. Therefore, it would seem appropriate to begin by evaluating his success at this endeavor.

My statistical consultant and I agree that Saunders' critique of Hyman's factor analyses is appropriate. Although Hyman has labeled his analyses as "exploratory," they were nonetheless used as an important part of his argument and thus demand critical attention. Strictly speaking, the analyses do not support Hyman's conclusion, culminating in his regression analyses, that a linear combination of three flaws--randomization, feedback, and documentation--can account for the significance of the results.

However, there is reason to qualify somewhat this harsh assessment. In particular, Hyman noted--regrettably, without providing supporting statistics--that four of the six procedural flaws (including those related to multiple analysis) were independently and significantly correlated with the primary (or, at least, the most sensitive) measure of statistical significance, the Z-score. Even allowing for multiple indices--a criticism which Hyman was quick to level at the parapsychologists but to ignore in his own work--I find this outcome noteworthy. The three most pervasive of these flaws are incorporated in the "Cluster III" (the "controls" cluster) of his first factor analysis, which, irrespective of the validity of the factor analysis, can be viewed as a kind of composite flaw scale.

Because of the questionable nature of the factor analysis, I have chosen to proceed by treating the components of Cluster III separately and to explore in each case the validity of the claim that the flaw is significantly related to the Z-scores. If these prove not to hold up, the factor analyses are invalid in any case. I have chosen Hyman's Z-scores as the dependent variable rather than Honorton's, because they are provided for more of the studies than the more conservative exact-probability Z-scores supplied by Honorton and are highly correlated with the latter (r=.99). I have chosen to express the relationships between the flaws and these Z-scores as biserial correlations, which provide an index of the magnitude of relationship as well as the significance. On average, these tend to give slightly higher correlation values (favoring Hyman) than the point-biserial correlation, which might also be justified. My criterion for significance will be a generous $p<.05$, one-tailed.

For each of the flaws to be considered, Honorton provided correlations with Z-scores which are nonsignificant, thus contradicting Hyman. There are two major differences in procedure that must be considered in evaluating these discrepencies: (1) Honorton and Hyman differed in their flaw codings of several studies and (2) Honorton restricted his correlations to the "direct-hits" experiments which comprised 28 of the 36 studies for which Z-scores were available (78%). Not only does this mean that Honorton's correlations were less powerful than Hyman's but, given the sizable mean difference of Z-scores between the direct-hit experiments and other experiments, Honorton's correlations may have been attenuated due to a restriction of range on the dependent variable.

Randomization. Hyman used three categories of randomization codings: (1) appropriate, (2) inappropriate, and (3) inadequately described. He claimed that the results were the same whether group (3) was combined with group (2) or omitted. Since combining groups theoretically provides the

most powerful analysis and appeared to be Hymans' primary method, I began using this approach. The resulting biserial correlation, based on the full sample of 36 cases, was +.372 (Z=1.65, p=.05), thus confirming (barely) Hyman's conclusion from his factor analyses.

There were 11 cases of randomization which Honorton coded as appropriate which Hyman coded as either inadequate or inadequately described. In 10 of the 11 cases I was able to find unequivocal evidence from the experimental report consulted by both reviewers that Hyman's minimal criterion of adequacy, "using a table of random numbers or a random number generator to select the specific target from a pool" (p. 27) was met. Recoding of these cases reduced the biserial r to +.02 (Z=+.08, n.s.). Thus, using this breakdown, Hyman's significant result can be entirely accounted for by improper coding of ten of the experiments.

However, this negative conclusion must be qualified on the basis of the analysis in which the experiments in Hyman's third category (inadequate description) are eliminated. Removal of these seven experiments causes the biserial r to rise substantially to +.385 (Z=1.87, p<.05). The result is essentially the same when the analysis is confined to the direct-hit studies (three of which fit in the third category, leaving N=25): r=+.407 (Z=1.84, p<.05). Thus, even with corrected codings there is a marginally significant difference between studies clearly using "adequate" and "inadequate" randomization procedures, the latter category yielding the higher mean Z-score.

This result was not reflected in Honorton's analyses because most of the studies judged inadequate by Hyman used a shuffling procedure, to which studies Honorton assigned the intermediate coding on his three-point scale. Without commenting at this point on the reviewers' disagreement about the absolute merit of the shuffling technique, I must disagree with Honorton's coding of these studies as higher than those where the method was not described (usually because the report only appeared as an abstract).

Feedback. Only the first of Hyman's feedback criteria (i.e., randomization of the judging sets) can be evaluated.[1] Since the feedback flaw so defined applied only to the 21 experiments which employed single target sets, the remaining cases are eliminated from the following biserial correlations. The correlation for the surviving 21 cases, using Hyman's codings, was +.565 (Z=2.76, p<.01).

Honorton noted two cases assigned the feedback flaw by Hyman in which it was indicated in the report that the members of the target set were rearranged in an order independent of the original target locations, descriptions comparable to those in other studies to which Hyman did not assign the flaw. (In one of these two cases [Sondow, 1979], the method of reordering was only specified for one of the two experimental conditions, but this was the condition responsible for the significance of the study and it is reasonable to infer that the same method of reordering was used in the other condition.) Honorton also cited two cases where Hyman did not assign a flaw but should have done so. My own inspection of the reports confirms Honorton in all these cases. Moreover, I uncovered two cases among the eight "non-direct-hit" studies excluded from Honorton's analysis in which the flaw was not assigned by Hyman but should have been (Habel, 1976; Parker, 1975). When the coding errors are eliminated, the biserial correlation is reduced to +.141 (Z=0.42, n.s.).

In conclusion, there is no significant relation between presence of the feedback flaw (first criterion) and significance of outcome in this data base.

Documentation. Hyman defined this flaw in such a way that makes independent evaluation of its coding extremely difficult. The one criterion he specified, which is the major one but apparently not the only one, concerns "the number of times the agent was a friend of the percipient or to

provide data on whether this made a difference..." (p. 28). First, it is never stated (to my knowledge) in any of the reports that all the percipients were unknown to the agents, even in those cases (generally not assigned this flaw by Hyman) in which the agent was one of the experimenters and the percipients were college students. Second, failure to evaluate this variable does not by itself constitute a flaw any more than failure to evaluate the sex, race, or hair color of the subjects constitutes a flaw. It is only a flaw if one specifies why this particular variable should have been evaluated, and Hyman was silent on this point.

Reading between the lines, one can make a highly plausible inference about what is going on here. Throughout his paper Hyman scrupulously avoided any mention of possible fraud by either subjects or experimenters, although we know that critics traditionally are obsessed with this problem. If one assumes that some subjects were motivated to cheat and (sensory) communication between agent and percipient is one of the most obvious ways that such cheating might occur, then cheating is a more likely possibility on those trials where agent and percipient were friends and relatively likely to be in collusion. Indeed, communication from percipient and agent, allowing the agent (in some cases) to reselect the target to bring it more in line with the percipient's mentation report, was the second criterion for the feedback flaw, which Hyman seems to have abandoned.

Concern with possible fraud by the agent might also have influenced Hyman's selection of the first feedback criterion (randomization of the judging set). In this case, I suspect he had in mind the first of the ganzfeld experiments (Honorton & Harper, 1974), in which the agent, who was sometimes a friend brought by the percipient, was responsible for reordering the judging set prior to judging by the percipient. If agent and percipient had been in collusion and knew something of the procedure beforehand, it would have been a fairly simple matter for them to have agreed that the real target would be placed in a certain location in the judging set. (However,

this would have been risky if the percipient's mentation obviously favored one of the other targets.) Assuming the experimenters were honest, this possibility only existed for those trials where the percipient brought a friend to be the agent, and it could be tested by reporting those trials separately. This was not done, hence assignment of the "documentation" flaw. This particular scenario--final reordering of the judging set by a subject--does not appear to have been possible in any of the other experiments in the data base.

In other words, Hyman's documentation flaw seems to be an indirect way of criticizing parapsychologists for a lack of sensitivity to the possibility of subject fraud in their experiments. However, if this is the case, it seems unfair to condemn an experiment on these grounds without considering what precautions were actually taken to prevent agent-percipient communication by sensory means.

Honorton seemed to have had the same hypothesis in mind when he attempted to operationally define the flaw by assigning it to studies, and only to studies, where (1) there was an agent, (2) the agent was not always the experimenter, and (3) when the second criterion was not met, there was no analysis of the effect of agent-percipient relationship on scoring. If one is willing to assume honesty on the part of the experimenters (which both Honorton and Hyman seemed to be doing), Honorton's criteria seem reasonable.

According to Hyman's codings of all 36 cases, the biserial correlation with the $\underline{Z}$-scores was +.473 ($\underline{Z}$=2.71, $\underline{p}$<.01), confirming the result of his factor analysis. Restricting himself to the "direct-hit" experiments, Honorton, using <u>his</u> criteria, uncovered ten cases in which Hyman assigned the flaw but should not have. My own survey of the reports caused me to concur with Honorton in all cases. Also, among the remaining experiments, I found one (Parker, 1975) which met Honorton's "flaw" criteria but was not assigned a flaw by Hyman. (In both cases, I could understand how Hyman

might justify his codings by his criteria.) Incorporating all the coding changes in line with Honorton's criteria reduces the biserial correlation to +.008 (Z=0.04, n.s.). Retaining all Hyman's codings for the non-direct-hit studies raises this correlation only slightly to +.178 (Z=+0.60, n.s.).

In conclusion, to the extent that Hyman's documentation flaw can be objectively defined by an independent analyst it is not significantly related to study outcome. To the extent that it cannot be so defined, the reader should not be asked to take it seriously.

Statistics. Using Hyman's codings, which were not disputed by Honorton, the biserial correlation with Z-scores was +.183 (Z=+0.82, n.s.), which fails to support the significance relationship reported by Hyman.

In conclusion, this evaluation has clearly failed to support Hyman's claim of a significant relationship between the presence of procedural flaws and the statistical significance (Z-scores) of the experiments in the data base in three of the four cases. Only in the case of randomization could some supportive evidence be found, but the level of significance was marginal and it was selected from among two equally plausible significance tests of the relationship. Thus, an appropriate correction for the duplicate analysis would render it nonsignificant. Moreover, even if it were left to stand, it would now constitute the only significant case among nine procedural flaws considered (including the three multiple-analysis flaws incorporated in Hyman's first factor analysis). In short, there is no credible positive evidence in support of a relationship between the flaws considered by Hyman and the outcomes of the experiments in the data base.

This result places the burden fully on Hyman's argument that the presence of the flaws constitutes sufficient cause in itself to reject the experiments in the data base as evidence for psi. I agree with Hyman's point that absence of a significant relationship with study outcome is no

guarantee that the flaw had no causal effect on the outcome. As he argued, the fact that each flaw is not independent of other flaws or procedural variables allows that suppressor variables might serve to obscure a real relationship. For example, the reason that the mean Z-score of experiments using single target sets did not significantly exceed the mean Z-score of the other experiments could be that the experimenters in the flawed category whose results were nonsignificant took more care to avoid the possibility that agents would contaminate the target pictures in the course of handling them than did those in the flawed category whose results were significant (perhaps due to the flaw). Likewise, demonstrating that all the studies in subsamples defined by the lack of particular flaws remain collectively significant is not an adequate rebuttal, because it implicitly assumes that only one flaw is operative in the data base. However, Honorton did not seem to be denying that flaws could account for the results. The real question is what kind of a case can be made for the proposition that they did account for the results.

To answer this question, we must first consider the plausibility of the mechanisms of the causal agents implied by the flaws under consideration. What other assumptions must be granted in order to conclude that these flaws accounted for the results? Let us consider the flaws individually.

Single Targets. If an agent and percipient had colluded to produce a bogus hit in any of the experiments cited for this flaw, it would have been a relatively simple matter for the agent to introduce cues onto the target picture detectable by the percipient. Since both Honorton and Hyman seem to have assumed honesty on the part of the experimenters, this possibility is really relevant only to those trials where the percipients brought their own agents. As noted, this seems to be the root cause of Hyman's concern that separate information was not provided about the results of such percipients.

If the agent did not intentionally introduce cues, is it possible that a percipient might still identify cues introduced unintentionally through normal handling? In an experiment conducted by the reviewer designed to answer this question (Palmer & Kramer, in press), it was found that "percipients" were indeed able to detect fingerprints left on the target photographs by "agents" who simply held them and lightly traced the index finger over them.

This criticism does not apply to those studies which used slides as targets. Palmer and Kramer found no support for a related suggestion that percipients are able to detect the slide previously exposed to the agent because it might have been more faded than the control slides. Moreover, the four studies which used slides and to which this criticism might apply collectively produced a Z-score less than zero.

Secondly, caution should be exercised in generalizing the Palmer-Kramer results to the procedure used in nine ganzfeld experiments where the targets were View-Master slides. Although fingerprints could have been introduced inserting or removing the slide from the projector, agent handling was less in these experiments than in those involving pictures, because the slide was being viewed while inside the projector. Because the projector uses only ambient room light, heat cues are quite unlikely. Finally, the judge evaluates the slides primarily while they are inside the projector, providing less opportunity for perception of fingerprints, etc. In short, creation and detection of sensory cues would seem somewhat less likely with this paradigm than with the one tested by Palmer and Kramer, although direct empirical evidence to this effect obviously would be desirable.

Returning to the experiments using photographs as targets, it should be noted that the fingerprints in the Palmer-Kramer study were not so obvious that they would draw the attention of someone not looking for them. In other words, it does not seem likely that a percipient not specifically looking for such cues would be biased by them. This conclusion is

reinforced by an earlier free-response ESP experiment by Palmer (1983), who systematically manipulated whether or not the photograph handled by the agent was included in the judging set. In most cases the agents and percipients were friends who shared an interest in parapsychology. No significant results were obtained in either condition, even though percipients (blind to the manipulation) were encouraged to look for "psychic vibrations" on the photographs in the judging sets.

In conclusion, the single-target hypothesis seems most viable in cases where the percipients are (1) aware of the possibility of detecting the target by means of sensory cues and (2) exert some effort to look for them. In cases where the percipients are "psychics" out to make a reputation for themselves or, at the other extreme, college students uninterested in the subject matter but wanting to help the experimenter "make the experiment work," the hypothesis seems quite plausible. However, in most of the experiments in the data base, especially those which were cited for the flaw and obtained significant results, the subjects seemed to be more like those in the Palmer (1983) experiment: volunteers who were oriented to assessing their psychic talents. For this kind of subject population the hypothesis seems less plausible, although it is still possible that such subjects might respond to subliminal cues or for some other reason take advantage of cues consciously perceived. However, the data from Palmer (1983) do not support such speculations.

Feedback. Failing to randomize the order of pictures in the judging set could bias the results if the target consistently appeared in one location that corresponded to the response biases of the subjects. For instance, in my experience subjects tend to have a bias favoring the first picture they see. If the targets were consistently placed first in the judging set, a spurious rate of excess hits could result.

This flaw was cited in those cases where randomization was achieved by hand shuffling of the members of the target set (these all turned out to be View-Master slides) or the method was not reported. The adequacy of the former depends on the care with which the shuffling was done. The latter is really a documentation flaw and cannot be evaluated.

Another possible way in which the feedback flaw could manifest in studies with the single-target flaw, which was not considered by Hyman, is replacement of the target in the set in a different orientation (i.e., upside down) from the other pictures in the set. Only a handful of the reports stated explicitly that this possibility was avoided.

Documentation. Failure to break down the results in terms of the relation of the agent to the percipient is only a flaw if there is reason to believe that it would make a difference, and in this case it only makes a difference if the possibility of fradulent agent-percipient communication has not been eliminated. Thus, this documentation flaw is really a by-produce of the sensory-cue and security flaws and need not be treated separately for present purposes. In other words, if one grants proper security against sensory communication, failure to report the breakdown at issue is at worst a minor transgression.

Security. This category cannot be evaluated for plausibility because Hyman never proposed how he thought certain failures to constantly monitor the agent and/or percipient could have allowed cheating to occur. Absence of such monitoring does not in and of itself mean that security was lacking. Honorton cited one experiment (Braud et al., 1975) where considerable care was taken to avoid sensory cues despite the involvement of only one experimenter, which was one of the criteria Hyman used to assign the flaw.

Statistics. Since Hyman and Honorton both computed their own statistics, which were essentially equivalent, the statistical indiscretions of some of the original authors do not affect the analyses at issue.

Randomization. Apart from incomplete or inadequate descriptions of the randomization methods in many of the reports, the crux of the issue here is Hyman's disapproval of a method used in one of the most successful laboratories, which involved selecting the target by shuffling a deck of 31 cards. Shuffling is not necessarily an inadequate method; Epstein (1977), for example, noted that six dovetail shuffles can adequately randomize a deck of cards. Thus, the method is only inadequate if one assumes that the shuffling was not done thoroughly. Again, the real flaw may be inadequate description of how thoroughly the shuffling was carried out. The very same point applies to the RNT and REG methods which Hyman found more acceptable. For instance, we saw in the discussion of the Maimonides dream experiments how improper use of a random number table can lead to biases that are likely to be more severe than those caused by casual shuffling. The point here is that the method of randomization is less important than how it is implemented. The fact remains, however, that it is easier to document proper application of REG and RNT methods than of shuffling methods, and in that sense the former methods are clearly superior. Unfortunately, as Hyman notes, the precise methods of randomization were rarely described even in those ganzfeld studies which used REGs or RNTs; this general deficiency of documentation is really his most potent criticism.

I cannot agree with Honorton's claim that the potential bias due to (inadequate) shuffling is minimized in cases where the randomization is performed separately for each trial. If shuffling is inadequate, the composition of the deck may not change sufficiently from trial to trial, thus allowing certain cards to consistently remain near the top of the pile and thereby have a greater chance than other cards to be selected repeatedly

during the course of the experiment. This could result in one of the targets being selected more frequently than chance allows. If this happens to be a picture that corresponds more closely to spontaneous mentation in the ganzfeld (discounting psi) than do others, a spurious level of hitting could result.

Nevertheless, the bias would have to be fairly extreme to compromise the results of experiments with the small number of trials characteristic of this data base. The shuffling-bias hypothesis is particularly strained in three experiments (mean $\underline{Z}$=1.23), where two or more cards were used to select a slide from the binary target pool of 1024 slides, which itself was ordered in a very complicated and nonlinear manner with respect to content (Smith, Tremmel, & Honorton, 1976; Terry, 1976; Terry, Tremmel, Kelly, Harper, & Barker, 1976).

It also should be kept in mind that <u>any</u> randomization method will occasionally produce ordered sequences that will correspond with subjects' response biases and thus leave the potential for spurious rates of hitting. This of course is the kind of thing reflected in the specification of the Type I error rate and is allowed for in the kinds of meta-analyses conducted by Honorton and Hyman.

Conclusion

The purpose of the above exercise was not to excuse the flaws cited by Hyman. In most cases they could have and should have been avoided and the experimenters deserve to be criticized for them. Nonetheless, a reviewer who is assigned the task of interpreting the significant results in the data base must be concerned with the question of how serious the flaws are, how likely they are to account for the results.

In most cases the flaws are attributable wholly or in part to inadequate documentation in the reports. In fact, the only flaw which clearly cannot be attributed to inadequate documentation was the use of

single target sets.

It should be kept in mind that the objective of the ganzfeld research was not to satisfy critics about the existence of psi but to explore a technique for enhancing the reliability of ESP in the laboratory. Most were written with the assumption that the primary audience would be other parapsychologists who would be interested in the studies primarily from the point of view of the objective which guided them and willing to take for granted that their colleagues would exercise basic experimental competence. This at least partly explains why the reports lacked details of interest to the critic, whose primary concern is whether any evidence for ESP exists at all.

With respect to duplicate target sets, it should be borne in mind that time and expense are often involved in creating such duplicate sets. This is particularly true in the case of the pool of 1024 specifically designed "binary target pool" slides (Honorton, 1975) used in several experiments; only a few of such pools were in existence. In the early days of the ganzfeld, and given the objectives of the research, it is understandable that researchers would want to forego that time and expense until the reliability of the procedure had been better established, especially since at the time it was reasonable to conclude that the possibility of biasing the results by handling cues was remote.

In retrospect, this may not have been a wise decision. Nonetheless, given the considerations discussed above, this and the other "flaws" Hyman uncovered do not seem to be of the nature that would justify an inference of general sloppiness in the conduct of the experiments.

In the final analysis, Hyman has failed to make a case that the flaws he uncovered provide an adequate explanation for the significant results in the data base. First, his internal analyses of the data have failed to provide any credible positive evidence for a causal relation between the flaws and study outcome. Second, he did not even attempt to argue for the

plausibility of the "normal" hypotheses suggested by the flaws, and my own analysis suggests that--assuming reasonable competence and honesty on the part of the experimenters--the hypotheses are not particularly plausible. On the other hand, the possibility that collectively these hypotheses can account for the results cannot be ruled out. To show this seems to have been Hyman's minimal objective and to that extent he has succeeded.

In conclusion, the ganzfeld experiments offer a genuine anomaly for which no adequate explanation exists. The explanation can only be obtained by further research.[2]

NOTES

[1] There is strong circumstantial evidence that Hyman did not in fact utilize his second feedback criterion (agent-percipient communication before feedback) since: (1) the second criterion was not mentioned in the discussion of this flaw in the text of Hyman's paper; (2) all experiments to which Hyman assigned this flaw used single target sets, a precondition for the first criterion but not the second; (3) there were only two experiments to which Hyman assigned the flaw and Honorton (who only coded the first criterion) did not, and in one of these it is easy to see how a miscoding vis-a-vis the first criterion could have occurred.

[2] Brief commentaries by other researchers on the ganfeld debate, as well as a reply by Hyman to Honorton's defense, are scheduled to appear in a forthcoming issue of the Journal of Parapsychology.

Chapter 5

RANDOM-EVENT-GENERATOR RESEARCH

I. Schmidt Research Program

Perhaps the most important methodological advance in experimental parapsychology during the past 15 years has been the introduction of electronic random event generators (REGs), also called random number generators (RNGs), for testing restricted-choice ESP and, more predominantly, PK. Although the REG was first used in psi research in the 1930s (Tyrrell, 1936) its emergence as a standard piece of apparatus in the psi laboratory can be traced to the work of physicist Helmut Schmidt (1969b, 1970c).

Schmidt's device is representative of the REGs presently being used in psi research. Electrical pulses pass a gate and arrive at a rapid rate (e.g., one million per second) at a switch, which advances one step each time. The switch is periodically stopped at one of its locations. To introduce a random element into this system, the choice of this location is influenced by a time delay based on the arrival and registration of a beta particle from a radioactive source (strontium-90) at a Geiger-Müller tube or by the peaking of the output of an electronic noise source. The location at which the switch stops defines the target selected by the REG. This selection is recorded on mechanical counters inside the machine and in some cases is also registered on paper-punch tape for a permanent record. In other applications, the REG can be interfaced to a computer and the results recorded on disk or tape. The selection is also fed back to the subject by the lighting of a lamp on the machine's face or by means of the attachment of the machine to a peripheral output device. This description will be further elaborated momentarily when particular applications are discussed.

Although the use of REGs has become widespread in parapsychology, by far the strongest and most consistent results have been from Schmidt's own

research program. Therefore, the first section of this chapter will be devoted to a review of this program.

There are certain characteristic features of Schmidt's program that are worthy of mention at the outset. Whereas most REGs are either boards that fit inside microcomputers (especially the Apple II) or are otherwise bound to main-frame computers or other laboratory hardware, Schmidt's machines are self-contained and portable. This allows them to be used "in the field" (e.g., subjects' homes), which provides for a more relaxed and informal testing environment. Schmidt feels this is crucial for obtaining positive results. He also places a great deal of emphasis on properly motivating his subjects and likes to have them "play" with the machine (i.e., do practice tests) prior to formal tests, often undertaking formal tests only at times when the subject is doing well on the practice tests.

Schmidt has published 14 experimental reports dating from 1969 to the present. The methodology has evolved through four somewhat distinct phases.

Phase 1. In this phase, when the emphasis was primarily on testing for ESP, the subject initiated each trial by pressing one of four buttons each located under a lamp on the machine. This button-press caused the machine to randomly select one of its four states as a target, which was indicated to the subject by the lighting one of the lamps. In the first experiment (Schmidt, 1969b), this was presented to the subject as a precognition task, i.e., the subject was asked to guess which option the machine would select and indicate his guess by pressing the button underneath the lamp of choice. However, Schmidt recognized that the subject could also be successful either by pressing the button at just the right time (precognition) or by forcing the machine by PK to select the option guessed. In a subsequent experiment (Schmidt & Pantas, 1972), Schmidt attempted to distinguish these possibilities (at least to a degree) by building into the machine an option such that, no matter which button the subject actually pressed, the subject

got a hit only when the target "4" was selected by the machine. This mechanism precluded simple precognition as a possible explanation of a significant result. The feedback was set up such that subjects were blind as to whether the machine was functioning in this "PK" mode or the standard precognition mode. Scores were significantly above chance to about the same degree in both modes. Simple PK was ruled out in another version of the experiment that was designed to test for clairvoyance (Schmidt, 1969a). In this case a sequence of targets derived from a random number table and prepared by an outsider was fed into the machine and substituted for the machine-generated targets.

Phase 2. In this and subsequent phases, Schmidt's focus shifted exclusively to PK. Instead of the subject activating each single trial by a button press, a whole series of trials (or a run) was so initiated. The number of trials per run varied from 100 to 1000, and the rate of event generation varied from 1 to 300 per second. Runs were short, lasting from a minimum of a few seconds to a few minutes each. The number of target alternatives were generally two (i.e., binary) rather than four as in the previous phase, although in later years probabilities as low as 1/64 were utilized. Different modes of feedback were also explored. In the first experiment of this phase (Schmidt, 1970b), the lamps on the machine were arranged in a circular display and each hit or miss caused the light to advance in the "right" or "wrong" direction around the display in a kind of "random walk". Auditory feedback such as clicks or variable-frequency tones were also utilized, as well as continuous polygraph tracings (e.g., Schmidt, 1973). In some cases where fast event generation rates were used, the feedback was integrated over blocks of trials. In two experiments, the subjects were animals and the feedback was selected so as to have reinforcing properties. In one case, the subject was a cat placed in a cold (0 degrees Celsius) environment, and generation of the target events caused

a heat lamp to be turned on (Schmidt, 1970a). In another case, generation of target events allowed cockroaches to avoid electric shocks (Schmidt, 1970a, 1979a).

Phase 3. In the mid-1970s, Schmidt published a mathematical model of psi inspired by physicists' attempts to interpret the "measurement problem" in quantum mechanics (Schmidt, 1975). From an experimental point of view, the most important implication of the model was that the choice of which alternative the machine selects on a given trial is not determined at the time of event generation but rather at the time the outcome of the event is "measured" (observed) by a conscious entity, i.e., at the time of feedback (Schmidt, 1976). If the observer happens to be a "psi source," the probabilities of the occurrence of the possible outcomes diverge from their a priori expected values in a manner related to the intent or wishes of the observer, thereby producing significant evidence of psi. Schmidt's model is an example of the so-called "Observational Theories."

In the methodologies of the preceding phases, recording of each event on the internal counters of the machine and feedback of that event to the subject occurred virtually simultaneously. The new model inspired Schmidt to develop a methodology in which these procedures are separated in time by minutes or even days. The most common procedure was to record a sequence of events on magnetic or paper tape and later to have a subject listen to or observe the tape or some other representation of its content. In most cases, the tape was presented in such a way that the subjects were led to believe they were receiving immediate, contemporaneous feedback as in an ordinary PK task (e.g., a Schmidt machine was used with the feedback being defined by the tape). Usually, the subjects were asked to attempt to influence the feedback (e.g., increase the number of randomly produced clicks), although in a few cases they were asked to merely observe it. In one experiment, a 1/64 hit rate was used and the subject had to wait for a

click (i.e., a hit), which terminated the run (Schmidt, 1976, [Exp. 1]).

Results with these "pre-recorded" events were compared with results of control conditions consisting either of contemporaneously generated events or of pre-recorded events which were statistically evaluated but never observed on a trial-by-trial basis. In one experiment, for example, pairs of sequences were generated for each run and one member of each pair was randomly selected to be presented to the subject as feedback. Only the observed sequences proved to be statistically biased (Schmidt, 1976, [Exp. 1, confirmation series]).

Phase 4. In this most recent phase, the focus on PK with pre-recorded events has continued. However, instead of each trial being generated randomly and therefore susceptible to PK influence, true random selection is limited to a "seed number" consisting of just a few digits. The pseudo-random sequence of events fed back to the subject is generated from the seed number by an algorithm which is impervious to PK (Schmidt, 1981). The point seems to be that the seed number is selected so as to produce a statistically biased sequence of events even though it is not directly observed by the subjects. However, the theoretical rationale for the procedure has not been fully developed or at least not fully articulated.

Statistical Evaluation

The output of Schmidt's REGs is amenable to simple and straightforward statistical evaluation by normal Z-tests, or critical ratios (CRs) as they are called in parapsychology. Schmidt has stuck to such tests almost exclusively, although slight modifications have occasionally been necessary, as in cases where results of machines having different a priori hit probabilities are to be pooled (e.g., Schmidt, 1976, [Exp. 3]).

Main Results

There is virtually no doubt that the results of the Schmidt REG experiments cannot be attributed to chance. The 14 articles I reviewed cited results of 33 experimental series, including 10 labeled "preliminary," "pilot," or "exploratory." Based on Z-tests of the total pooled trials in each series, 25 of the 33 (76%) were significant at the .05 level, two-tailed. In two of the seven nonsignificant studies, one of the two experimental conditions yielded significant results (Schmidt, 1979b; Terry & Schmidt, 1978). Of the remaining clearly nonsignificant series, four involved tests on sub-mammalian species (e.g., algae, fruit flies) where it is not clear whether significant results were expected (Schmidt, 1979a). In one of the others, which also involved an infra-human species (cat), significant scoring occurred in the first half of the test and the cat seemed to have been noticeably less oriented toward the reward stimulus during the nonsignificant second half (Schmidt, 1970a).

Of the 29 series where the direction of the results could be determined from the report, scoring was in the direction presumably intended by the subject in 21 of them (72%). When the direction of the Z-scores was defined relative to the subject's intent (which was assumed to be for hitting unless specified otherwise) and cumulated across series by the Stouffer method (Mosteller & Bush, 1954), the resuling Z was a whopping 9.92.[1] This figure is somewhat conservative, because in four cases psi-missing was predicted by Schmidt although presumably not intended by the subjects (Schmidt, 1970a, confirmatory series; 1970b, confirmatory series, Schmidt & Pantas, 1972, both series with groups). When the direction of the individual Z-scores was defined in terms of Schmidt's predictions, the cumulative Stouffer Z was 14.85.

Because the a priori probability of a hit varies so much from experiment to experiment, it is difficult to provide a comprehensive estimate of the average magnitude of the scoring in Schmidt's experiment.

However, estimates can be provided for the ESP studies (Schmidt, 1969a,b; Schmidt & Pantas, 1972), all of which used P=1/4, and all but one of the PK studies where P=1/2 was used (Schmidt, 1970a [cockroach experiment], 1970b, 1973, 1974, 1976, [Exps. 2 and 3], 1978). (The cat experiment [Schmidt, 1970a] had to be eliminated because full data were reported for only half the trials.) Using the trial as the unit of analysis, the mean percentage of hits for the ESP studies (MCE=25%) was 26.50%; for the PK studies (MCE=50%), the mean was 50.53%. However, two-thirds of the trials in the PK sample came from one experiment, which produced an abnormally low hit rate, possibly because of an unusually rapid rate of event generation. With these trials removed, the mean is elevated to 51.26%, which I think is a more representative figure.

Results: Independent Variables

As a general rule the rather modest variations of experimental procedure that Schmidt employed in his experiments did not significantly affect the results.

Subjects. In his early ESP experiments, Schmidt (1969a,b) gave a large number of trials to a handful of subjects who had clearly shown promising results on screening tests. Later on, larger samples of subjects were used for whom screening was either minimal or absent, and there was no noticeable dropoff in scoring rates. However, Schmidt did insist on providing his subjects some opportunity to play with the machine prior to testing and he attributed one of his unsuccessful series to the fact that in this particular case subjects were not given that opportunity (Schmidt, 1978).

Feedback. Neither the type of feedback display nor the rate of feedback has affected scoring in Schmidt's experiments, although fully controlled comparisons have not been undertaken.

Pre-recorded events. Results of series with pre-recorded events (including pre-recorded seed numbers) have been comparable to those in series with real-time events. This conclusion also applies to experiments where the two types of trials were compared in the same series (Schmidt, 1976 [Exp. 2], 1979b).

There is some evidence that the following two variables might affect scoring, but the evidence is equivocal.

Rate of event generation. In one PK experiment with a binary REG, Schmidt (1973) found that the scoring rate was significantly lower when the machine generated targets at 300 per second (50.4%) than at 30 per second (51.6%). However, this conclusion must be viewed cautiously, because subjects were not assigned randomly to the two conditions, some subjects participated in both conditions, and subjects were not always blind to the generation rate. In two subsequent experiments (Schmidt, 1974), results with a machine generating trials at 100 per second were compared to results with a machine generating trials at one per second. In these experiments, trials from the two machines were interspersed and subjects were blind as to which generator produced each trial. In the pilot experiment with four subjects, scoring was significantly higher with the slow REG than with the fast one (58% vs. 51.2%). In the confirmatory experiment with 35 subjects, the difference was in the same direction but not significant (55.3% vs. 53.8%). Two experiments with pre-recorded targets (Schmidt, 1976 [Exps. 2 and 3]) also used generation rates of 300 per second and although no systematic comparisons are possible, the mean hit rate on these (fast rate) trials (51.8%) is actually higher than the overall average (51.26%).

Intent. In one experiment (Terry & Schmidt, 1978), trials where subjects were asked to use PK to alter the REG output were compared, within subjects, to trials where they were not to attempt PK influence but merely to attend to the feedback tones. Significant scoring was restricted to the intentional PK condition, although results were unexpectedly in the missing direction and significant by only one of the two tests the authors used. No statistics comparing results in the two conditions were reported. One other experiment in which subjects were asked to listen to the feedback without attempting to influence the REG yielded clearly significant positive results (Schmidt, 1976 [Exp. 1]). Collectively, these results suggest that "intent," at least in the sense of conscious effort to influence the REG, is not necessary to achieve the predicted effect but might be facilitative.

Criticisms

The two major critics of Schmidt's experiments have been C.E.M. Hansel (1980) and Ray Hyman (1981). Their principle concerns have been inadequate security against fraud and inadequate randomization tests.

Security. Hansel, who only considered the work of Schmidt through 1970, focused his criticism on the fact that in most of these experiments the target was changed during the course of the experiment. In two of the three ESP series (Schmidt 1969a,b) subjects sometimes aimed for hits and sometimes for misses. In the PK experiments (Schmidt, 1970a,b) the target was changed halfway through the experiment. Hansel's concern was that the nonresettable electronic counter inside the REG did not take account of the change in target and actually recorded chance outcomes in the series where the changes occurred. Although the changes were recorded on the paper tape that constituted the other independent record and the two records matched, Hansel felt that this was not an adequate provision. Hansel was somewhat vague in his criticism (perhaps for legal reasons), but it would appear that his

concern was less with innocent recording errors than with the possibility that Schmidt could have cheated by improperly assigning the target direction after the data had been collected.

Hansel seemed reasonably satisfied that the machine precluded fraud by the subject (at least he did not raise this possibility explicitly), but Hyman expressed concern that Schmidt seemed to place too much trust on the machine to provide security and that "...subjects, for the most part, are unsupervised and unobserved" (p. 37).

Randomization tests. For Schmidt's experiments to be evidential of psi, it is generally considered necessary that the output of the REG be unbiased in the absence of attempted PK influence. Although acknowledging that Schmidt did indeed conduct randomization tests, both Hansel and Hyman were concerned that the tests were not conducted systematically. For the most part, Schmidt conducted long series of control trials periodically during the course of an experiment. Hyman in particular felt that such tests might be insensitive to short-term biases that might operate in actual experiments, where the runs consist of many fewer trials. For example, the machine might only be biased for the first few trials after it is turned on and this would not show up in long control sequences. The critics suggested that a better procedure would have been to collect runs in pairs, randomly assigning one member of the pair as the experimental run each time.

Hyman also criticized the fact that the randomization tests did not check for biases beyond the second (doublet) level.

Evaluation

Security. Although the use of a machine like Schmidt's REG does minimize opportunities for subject fraud, it is nonetheless reasonable to insist that subjects not be allowed access to the machine unattended. However, Hyman's positive assertion that most of Schmidt's subjects were

unsupervised and unobserved is incorrect. Although the reports tend not to be as thorough on this point as one would like, it is clear that Schmidt is either with the subject or with the REG in most cases. In one case Schmidt (1969b) noted an exception for one subject but also noted that the results of the experiment did not depend on that subject.

The possibility of subject fraud is further minimized in the experiments using pre-recorded targets, as the subject is not present when the target sequence is generated. Subject fraud is remote indeed if an independent record of the target sequence is kept elsewhere while the targets are being observed by the subject. Such records were reported as being kept in experiments where the subject was allowed to take the tape home (Schmidt, 1976 [Expt. 1], Schmidt, 1978), but it is not clear whether they were kept in the other experiments. If independent records were not kept, it is conceivable that a subject might somehow substitute a biased tape for the one in which the target sequences had been recorded. However, this would require a fairly elaborate ruse on the part of multiple subjects, many of whom were relatively unselected volunteers with no apparent investment in being certified as psychic.

Fraud by the experimenter is, of course, always more difficult to rule out, and the fact that Schmidt usually works without a co-experimenter makes the hypothesis particularly tempting to some critics. It is difficult to understand why Hansel harps on the nonresettable counters in this connection, however, because it would have been easy for Schmidt, who builds the machines, to tamper with the counters either before or after a session if he were inclined to cheat, even if he were using the set-up Hansel recommends. Moreover, Hansel's critique fails to note that the goal for each set of trials (high-aim or low-aim) is registered on the paper tape along with the events themselves.

Schmidt has recently taken the experimenter-fraud criticism to heart, however, and devised a procedure using two experimenters. One experimenter

is responsible for generating the (pre-recorded) event sequence while the second experimenter determines the random order of target directions across runs or determines which of the recorded tapes are to be observed by the subject and which are control tapes. Thus, neither experimenter by himself can artifactually produce significant results by generating a faulty sequence. Variations of this procedure have now been tried in two experiments (Terry & Schmidt, 1978; Schmidt, Morris, & Rudolph, in press), once with significant results. Of course, this procedure does not rule out collusion by the two experimenters and so far all the experimenters have been sympathetic to parapsychology.

Randomization tests. The critics are correct in pointing out that Schmidt's early randomization tests do not adequately exclude the possibility of short-term biases, at least those that might occur just after the REG is activated for a run. However, the argument is weakened by the fact that the critics have so far not been able to articulate a mechanism that would produce such a bias. Short-term biases that would occur intermittantly at other times would have to be consistent in direction to account for the results Schmidt found in his experiments, yet in that case they also would accumulate and thus be revealed in the randomness tests Schmidt did undertake.

The one piece of empirical evidence that can be cited against short-term bias of the former type as an explanation for Schmidt's results is the fact that the deviations covaried with changes in target. Ironically, it is the very procedure that Hansel criticized for security reasons that provides this evidence. For example, the bias hypothesis would have to explain why the direction of these biases would suddenly and consistently change when a subject started aiming for misses instead of hits in the ESP tests (Schmidt, 1969a,b) or when the experimenter changed the target direction in the early PK tests (Schmidt, 1970a,b). It is not clear

how the bias hypothesis can account for such covariations without taking on unparsimonious ancillary assumptions.

In Phase 3, Schmidt began using control procedures which conformed more closely to those demanded by his critics (Schmidt, 1976). In the first two experiments described in this paper, Schmidt reported control runs of the same length as the test runs. Furthermore, in two of the experiments from this phase (Schmidt, 1976 [Expt. 1], Terry & Schmidt, 1978), pre-recorded control tapes were made at the same time as the pre-recorded experimental tapes; which tape was which was determined randomly--a procedure very close to that suggested by the critics.

It should be mentioned, however, that Schmidt is not consistent in his reporting of randomization tests. Half of the 14 reports I reviewed did not report randomization tests and the descriptions in the others varied in degree of detail. However, I could find no relationship between the adequacy of the randomization test as reported and the significance of the results.

Finally, what about the possibility of dependencies between random events beyond the second level? Such dependencies would, strictly speaking, violate the assumptions of the statistical tests Schmidt used. However, if they were to have biased the results of Schmidt's experiments they would have had to manifest at the singlet level as well, and if that were the case a singlet bias would also have appeared in the randomization tests. If the higher level dependencies did not favor the target at the singlet level, they would also have tended to diminish the likelihood of significant results in the experiment proper. Thus, the possibility of higher level dependencies does not appear to provide a plausible explanation of Schmidt's findings.

Data selection. A possible criticism not made directly by other critics concerns the possibility that the apparent significance of Schmidt's results

is attributable to selective publication of positive findings. Schmidt often refers in his reports to exploratory testing with various outcomes. He acknowledges a preference for testing subjects formally only at times when they are doing well informally, which implies that sometimes ubjects do not do well informally. REG experiments are very economical timewise, and it is conceivable that Schmidt has accumulated vast amounts of nonsignificant and unreported data.

However, Schmidt's procedure is not a problem if certain subgroups of trials are specified beforehand as formal tests and reported irrespective of outcome. Barring outright dishonesty on Schmidt's part, this provision seems to have been met by the 14 experiments in the articles I reviewed which Schmidt defined as "confirmatory" or "main" experiments. The collective results of these experiments are actually better than the overall average (Stouffer $\underline{Z}$=15.60).

A related potential artifact concerns optional stopping, in particular not specifying in advance the number of trials in the experiment. I determined that in 15 of the series reviewed, it is stated or clearly implied in the report that the number of trials was stated in advance. In two of these cases, two alternatives were specified in advance but the degree of selection could not conceivably account for the strong results obtained (Schmidt, 1969a,b [Exp. 2]). In a third case, a range of from 55,000 to 70,000 trials was specified (Schmidt, 1969b [Exp. 1]). Here the possibility of the artifact being fatal is still remote, but nonetheless conceivable. Discounting this experiment, the remaining 14 experiments with clearly prespecified numbers of trials have a Stouffer $\underline{Z}$ of 8.42. It is likely that in most of the other experiments the number of trials was also prespecified even though it was not formally stated.

Some commentators have expressed concern over Schmidt's procedure of not prespecifying how many trials each subject contributed to the total, especially since in his early work he would sometimes stop testing a

high-scoring subject when scores began to decline. However, it seems clear that such a procedure is statistically acceptable as long as the total number of trials is prespecified.

Source of the effect. An interesting point of disagreement among those who at least tentatively interpret Schmidt's research as providing evidence for a paranormal principle is whether the source of the effect is really Schmidt's subjects or Schmidt himself. Several points can be made in favor of the latter hypothesis:

1. Although other investigators have achieved significant effects in REG research, no one has done so as consistently as Schmidt or has achieved such strong (by parapsychological standards) effects.

2. There are growing indications that directing effort toward achieving a psi effect is not necessary for the effect to occur. This was demonstrated in one of Schmidt's own experiments (Schmidt, 1976 [Exp. 1]) as well as in experiments by others (e.g, Palmer & Kramer, 1984; Stanford, Zenhausern, Taylor, & Dwyer, 1975). These experiments suggest that it is the need to achieve a certain outcome in contrast to the effort to achieve it that is crucial. If so, then Schmidt, as the experimenter desirous of positive results, becomes at least a potential psi source.

Schmidt's model assumes that observation (feedback) of the trials of an experiment is necessary for influencing them. In one of his later experiments, he indeed established that only those pre-recorded event sequences observed by his subjects (and not control tapes generated at the same time) were biased, thereby suggesting that the subjects biased the original generation of the sequences retroactively at the time of feedback (Schmidt, 1976 [Exp. 1]). However, coming from a more traditional theoretical orientation, one could argue that it was Schmidt who used PK proactively to bias the event sequences at the time they were being generated. While it is true that Schmidt did not know at the time of

generation which member of each pair of sequences would be the experimental one and which the control, this lack of knowledge does not necessarily preclude his having used psi, albeit without conscious intention to do so. Schmidt himself has been foremost among those parapsychologists arguing that psi is "goal directed" which implies, among other things, that one can achieve a desired goal by PK without knowing the mechanism needed to produce it. One of Schmidt's other experiments (Schmidt, 1974) has shown that significant results can occur without subjects' knowing which machine is operative on a given trial. Subjects in another experiment (Schmidt, 1981) did not realize that their real target was a random seed number and not the pseudo-random event sequence derived from it and to which their attention was directed; in fact, they never observed the seed numbers at all. Thus, it is questionable whether Schmidt's ignorance of the contingencies at the time he generated the pre-recorded event sequences is any greater or more important than the ignorance of the supposed subjects when they observed them.

3. In one experiment (Schmidt, 1970a) the subjects were cockroaches and the REG was biased so as to increase the number of painful shocks they received. At least from the standpoint of motivation theory, this result makes little sense if one assumes that the cockroaches were the psi sources. It makes somewhat more sense if one assumes that Schmidt was the psi source, especially if one is safe in assuming that Schmidt does not like cockroaches!

II. Princeton Research Program

The other major research program in parapsychology using REGs is being undertaken by Robert Jahn, Dean of the School of Engineering at Princeton University, in collaboration with psychologists Roger Nelson and Brenda Dunne (Nelson, Dunne, & Jahn, 1984). The two REGs they employ are similar in basic concept to Schmidt's, but no radioactive decay source is utilized.

Instead, the output of a commercial electronic noise source is filtered, amplified, and sampled by a train of gate pulses. The target is determined by the sign of the noise (above or below the zero crossing) at the time of each sampling.

The Jahn team uses different terminology than does Schmidt to define the sampling regimen. To maintain uniformity of exposition, I will translate Jahn's terminology into Schmidt's, as the latter is more standard within parapsychology.

In his formal tests, Jahn collected runs each consisting of 200 trials generated at either 100 or 1000 per second. Thus, Jahn's runs were much shorter than Schmidt's. Unlike Schmidt, Jahn alternates the target (from the point of view of the REG) between trials; i.e., positive and negative noise alternate as hits on successive samplings. The subject, whose task is to use PK to bias the REG output in the target direction, can activate the REG in one of two modes. In manual mode a button press activates only a single run. In automatic mode it activates a sequence of 50 runs. The subject receives continuous feedback on LEDs consisting of the number of runs so far completed, the number of hits in the last run, and the running average of run scores completed in the test. The latter is displayed most prominently.

The number of hits in each run is registered in the REG, which also computes and records the mean and standard deviation of the run scores of each 50-run block. These data are eventually transferred to magnetic tape for statistical analysis by computer. To provide redundancy, data are also recorded on paper tape and manually in a logbook.

Like Schmidt, Jahn tries to maximize the comfort of the subject and provides for practice runs prior to formal testing. Subjects are encouraged to undertake formal tests only when they are in the mood. Subjects can also determine the length of each session, provided it includes at least five blocks. Sessions are scheduled at the subject's convenience to the extent

possible.

Each session consists of some runs where the subject aims for a high score (PK+) and some runs where the subject aims for a low score (PK-). The sequence of target aims is sometimes determined by the subject's preference--volitional mode--and sometimes by means of an objective random process--instructed mode--the nature of which is not specified in the report. Interspersed among the PK runs are also a number of baseline runs initiated by the subject but for which the subject attempts no PK influence. These serve in effect as randomization tests.

The sessions were organized into series, which range from 500 to 7500 runs per condition. The formal experiments consisted of 61 series contributed by 22 subjects. Each subject contributed from 1 to 14 series, with 44% of the series contributed by two individuals. The total number of runs was 569,450 (or 113,890,000 trials). The latter is approximately 189 times the number of trials in all of Schmidt's published experiments combined.

Statistical analyses of the data consisted primarily of single-mean t-tests comparing mean run scores to MCE (=100), using an empirical variance estimate. This is in contrast to Schmidt, who uses Z-tests with a theoretical variance estimate.

Jahn also uses various graphical representations of his data, in particular, plots of the cumulative deviation of mean run scores over runs for PK+, PK-, and baseline runs, respectively. These are capable of reflecting variations in a subject's performance over time.

Results

The main presentation of results was restricted to those formal series consisting of 200 trials per run. These comprise 390,200 runs, excluding baseline runs. The mean number of hits per run for the PK+ runs (exactly half the total) was 100.04 ($t=2.68$, $p<.004$, one-tailed). The mean number of

hits for the PK- runs was 99.97 (t=2.16, p=.016, one-tailed). Defined in terms of the subject's intent (i.e., missing in PK- runs equals hitting), the combined results were associated with t=3.42, $p=3\times10^{-4}$, one-tailed. The percentage of hits was 50.02%. This is much lower than the 51.26%, or the more conservative 50.53%, estimated for Schmidt (see p. 103).

Jahn also broke down his data in terms of three independent variables: mode of activation (manual vs. automatic), mode of target-aim selection (instructed vs. volitional), and rate of event generation (100 vs. 1000 per second). These analyses revealed, first, that the significance was entirely attributable to runs where the target aim was selected voluntarily (volitional). The hit rate for these runs was 50.04% (t=4.24, $p<10^{-5}$, one-tailed) compared to 50.01% for the runs where the target aim was determined randomly.

Both event generation rates yielded significant overall results. However, the slower rate (100 per second) yielded slightly higher scores (50.05%) than did the faster rate (50.02%), especially in the PK- condition. I computed a t-test of the difference based on the data reported by the authors and determined that the superiority of the slower generation rate (for both PK+ and PK- runs combined) was significant (t=2.04, $p<.05$). However, this analysis may be misleading and will be discussed further in the evaluation section. Mode of activation had no significant effect on the results.

Examination of the cumulative run score graphs led the authors to conclude that their subjects had individual scoring patterns, which they called "signatures," that were consistent within each subject for a given specification of test parameters. However, no statistical analyses were offered to support this conclusion.

The mean of the 179,250 baseline runs was 100.005, which was acceptably close to the expected value of 100. The variance of the scores was also within chance limits, but a graph of the distribution exhibited a marked

excess of scores right at 100.

The authors later presented a separate set of analyses with the series as the unit (Jahn, Nelson, & Dunne, 1985). According to this analysis, the displacement of the mean Z-score remained significant for the PK+ runs (mean=+.377; $t[60]=2.59$, $p<.02$, one-tailed) but not for the PK- runs (mean=-.262, $t[60]=1.61$). As expected, the mean for the baseline was very close to chance (mean=+.023).

However, this analysis revealed some curious differences in variances. The variance of the series scores for the PK+ runs was suggestively high ($F[60, \infty]=1.295$, $p<.20$) whereas that for the PK- runs was significantly high ($F[60, \infty]=1.616$, $p<.02$). It is likely that the high variance was responsible for the failure of the mean to reach significance in the PK- condition. The really curious finding, however, was a corresponding restriction of variance of the series scores for the baseline runs that approached significance ($F[49, \infty]=.663$, $p<.10$). The restriction was caused by an absence of any Z-score values greater than +1.645 or less than -1.645 in the distribution. (The p-values reported here are more conservative than those reported by the authors, presumably because the latter were using one-tailed tests. I find two-tailed tests more appropriate for this application.)

In addition to the formal series, 34 exploratory series comprising a total of 103,950 test runs (PK+ or PK-) have been conducted (Nelson et al., 1984). Two subjects contributed all but one of the series. The procedure differed from that of the formal series in that the number of trials per run was 100 or 2000 instead of the usual 200, or the number of runs per condition per series was low (approximately 1000).

Only the cumulative results of five series with 2000 trials per run provided by one of the two subjects tested with this protocol (who happened to be the high-scoring subject in the formal series) yielded significant results, which were in line with the findings from the formal experiment.

Results from the remaining subject, who completed only one series at this rate, were close to chance.

Finally, 12 exploratory series totaling 60,000 test runs were conducted with the event sequence being generated from an algorithm as opposed to the REG. Such a pseudo-random sequence presumably cannot be influenced by a PK force, so success would appear to be possible only by entering the sequence at a point that would yield a "biased" subset of numbers embedded in it. Significant results in line with the formal series were nonetheless achieved in the cumulated seven series performed by the same subject who obtained the significant results in the previous series. The two subjects who completed the remaining series provided only chance results.

Baseline scores in the combined exploratory series, and specifically in each of those subsets which provided significant results in the experimental conditions, did not deviate significantly from chance expectation.

Criticisms

No critiques of Jahn's research by outside commentators have yet been published, to my knowledge.

Evaluation

Controls and security. The Jahn team has done a better job than Schmidt in providing adequate baseline tests, as Jahn's baselines were were all collected in the same sessions as the experimental data and were of identical structure. Jahn also reported more extensive tests of the machine's performance outside the test sessions, including checks on the function of separate components. Internal checks during all operations were used to assure that proper input voltage of the noise diode was maintained and that internal temperature was not correlated with machine performance. Recording of the run scores as well as preliminary data such as designation of run type was redundant and included automated recording.

The one deficiency I can cite regarding security is that Jahn does not fully document what precautions were taken to preclude data tampering by subjects. This omission is particularly significant because an experimenter was not present in the room with the subject during formal sessions. As it would appear that the fundamental hardware for testing, including the REG itself, was located in the same room as the subject, tampering with the equipment is a theoretical possibility. On the other hand, such tampering would seem to require computer sophistication on the part of the subject as well as (in all likelihood) knowledge of the particular setup being employed. If failsafes were utilized to preclude such tampering, they are not described in the reports.

Data selection. The Jahn team appears to have been conscientious in reporting all the exploratory and formal series they have undertaken. However, it is less clear that the distinction between these two subclasses of experiments was made in advance. The exploratory series yielded somewhat weaker effects overall than the formal series (50.01% hits vs. 50.02% in the formal series). If all the series reported were pooled, it is not certain that the overall result would differ significantly from chance.

However, this issue loses importance when one considers how the significance is distributed among the various subjects tested. In the formal series, only two of the 22 subjects tested provided independently significant results. The bulk of the significance is attributable to one of these subjects, who contributed 14 of the 61 formal series (23%). In these series, this subject achieved a hitting rate (in terms of his or her intent) of 50.05% over 105,150 runs ($Z=4.49$, $p<10^{-4}$). When the results of this subject are eliminated, the remaining series are no longer significant (50.01%, $Z=1.36$). This subject's scoring rate is significantly higher than that of the other subjects combined ($Z=3.14$, $p<.005$).

Moreover, this subject is the same subject who provided the significant results, and the only significant results, in the exploratory series. The overall results of this subject are clearly significant and consistently in the expected direction, i.e., above-chance scoring in the PK+ runs and below-chance scoring in the PK- runs. Twenty-eight of the 35 series (80%) in which this subject participated produced higher scores in the PK+ condition than in the PK- condition, which in itself is a significant outcome ($X^2[1]=12.60$, $p<.001$). The mean t of the 35 series is +0.84, s.d.=1.35, which is significantly different from zero ($t[34]=3.70$, $p<.001$). The one puzzling finding is the failure of this subject to maintain his or her usual scoring level in the short 1000-run series, which were otherwise methodologically identical to the formal series.

A more uniform distribution of scoring across subjects is suggested by analyses using the subject as the unit. A mean run score on the experimental runs was computed for each subject by reversing the direction of the PK- scores and taking the average of the PK+ and PK- scores, weighted by the number of runs in each condition. The mean of these scores was 100.03 which is significantly above chance, although barely so ($t[21]=1.74$, $p<.05$, one-tailed). However, when the experimental scores are contrasted to the baseline scores using a dependent t-test, the result falls just short of significance ($t[21]=1.67$). Neither analysis would be significant were the high-scoring subject eliminated, but this analysis nonetheless provides some evidence that the effect is uniformly distributed within the sample. However, the evidence is weak and can only be considered suggestive.

Optional stopping. It is not stated whether the total number of trials and the number of trials completed by each subject were specified in advance. In principle, this leaves the Jahn research open to the criticism that a series may have been terminated at times favorable to the support of the experimental hypotheses. An opportunity to test the optional-stopping

hypothesis is suggested by the fact that the degree to which optional stopping was potentially operative seems to have varied from series to series. In 24 of the series, the number of runs in the PK+ and PK- modes was identical. In 23 of these, the number of runs per type was either 2500 or 5000; in the remaining series it was 3000 runs. It is very unlikely that optional stopping was a factor in these series. Thus, if the optional-stopping hypothesis were correct, one would expect lower scoring in these series than in the other 37. As a dependent variable, I chose the *t*-score supplied by the authors which reflected the difference in scoring between the PK+ and PK- conditions for each series. The mean *t*-score for the 24 "uniform" series is +0.74, which is higher than the mean of +0.22 for the remaining series ($t[59]=-1.51$, n.s.) and opposite to the direction predicted by the optional-stopping hypothesis. In fact, the mean for the uniform series is independently significant ($t[23]=3.38$, $p<.005$). Thus, the optional-stopping hypothesis can be safely discounted.

Independent Variables

The interpretation of the results relating scoring levels to mode of target-aim selection and rate of event generation is complicated by the fact that not all subjects contributed equally to the various levels of these independent variables. Only three subjects contributed runs at the slow generation rate. Further analysis of the data reveals that the apparent superiority of scoring at the slower rate (see p. 115) is attributable to the fact that these three subjects had higher scores overall than the other subjects. (Note that the authors never claimed superiority for the slower rate, but it might be inferred from one of their tables.) Mode of target-aim selection (volitional vs. instructed) was more evenly balanced, but 12 of the 22 subjects contributed to only one of the two conditions. However, the superiority of the volitional mode of target-aim selection is so pronounced that the design confounds seem inadequate to account for it. The effect

held up in the data of the one significant subject but was actually stronger among the remaining subjects. In fact, the hit rate among the remaining subjects in the voluntary mode was 50.03% (Z=3.19, p<.01), providing further evidence that some of these other subjects may have exhibited some psi in the experiment.

Baseline

Although the statistical evidence for restricted variance in the baseline runs is in my judgment less than the authors claim, the suggestive trends that were uncovered are perhaps worthy of some tentative interpretation. Could it be, for example, that subjects unwittingly exerted a PK influence in order to assure that the baseline data were "good baselines," i.e., conformed closely to MCE? Such an interpretation would coincide with the assumption of Stanford's (1974a,b) PMIR model that PK does not require intention and effort on the part of the subject.

Intuitive Data Sorting

Up to this point, we have assumed that if the REG data are ultimately explainable by some paranormal principle, that principle implies some causal influence on the REG; i.e., it involves PK. An alternative interpretation is suggested by a model called "intuitive data sorting" (IDS) proposed by May, Radin, Hubbard, Humphrey, and Utts (1985) to account for REG PK data generally. According to this model, significance occurs because of a psi-mediated selection of the starting point of the sequence of random events so as to capture locally "biased" subsequences. For example, if significance is defined as p<.05, one of every 20 sequences from a truly random source should be significantly "biased." If such sequences were captured more frequently than 1 in 20 times, a cumulatively significant deviation could result. An attractive feature of the IDS model is that it seems to account better than competing causal models for the failure of

statistical significance to increase as N increases, a trend that is evident in the actual data base (May et al., 1985). It is also noteworthy that Schmidt had considered a similar hypothesis several years earlier (see p. 98).

May's model could be more appropriately labeled "intuitive data selection," since no sorting is thought to actually take place. In other words, favorable subsequences are selected from the total, ongoing sequence, but there is no preordained set of outcomes that are sorted into different categories. A true sorting model could, however, be applied to Jahn's data by hypothesizing psi-mediated assignment of "random" run scores to the PK+, PK-, or baseline categories. The fact that results were better in the "voluntary" mode than in the "instructed" mode could be interpreted as supporting such an interpretation, since the latter gives the subject more flexibility and control in selecting the run type. Such selection is possible in the "instructed" mode but it would require some kind of psi-mediated selection of the random number which is the direct cause of the determination of run type, and the decision would then be forced for the entire 50-run block.

An important implication of this model is that the total distribution of scores, irrespective of type, must conform to a true Gaussian. This criterion is met when the run is taken as the unit of analysis, but when the series is taken as the unit, there are not enough scores in the middle of the distribution to form a true Gaussian. This latter result, however, is not necessarily inconsistent with the sorting model. The expectation that the series scores should form a Gaussian distribution in this case assumes that the run scores are randomly assigned to type (PK+, PK-, or baseline) within the series. If this is not true, distortions of the distributions of series scores could easily result. For instance, the depressed variance of the series scores in the baseline condition could result from a tendency for the average ratio of above- to below-chance baseline runs within a series to

be closer to 50:50 than expected by chance. One could speculate, for example, that as subjects noted an unusually high proportion of outcomes in one direction in the baseline condition they began to produce outcomes in the opposite direction to compensate. High variance in the PK+ and PK- conditions could result if assignment of higher run scores to the PK+ condition occurred only in some of the series, an assumption which is consistent with the already stated conclusion that the anomalous scoring in these conditions seemed to be largely attributable to one of the subjects. Such a situation would cause a distribution of series t-scores which includes a small group of highly positive t-scores added to a larger group of t-scores closely conforming to a true Gaussian, thereby increasing the variance of the distribution as a whole. The important point is that the variance effects uncovered at the series level can be obtained by simple rearrangement of a perfectly random "chance" distribution of run scores.

Strictly speaking, the sorting model implies that the proper unit of analysis is not the run but whatever number of runs a subject completes in succession without having the option to change run type (e.g., PK+ to PK-). It is not stated in the reports what this unit is, and it may have varied from series to series or even within series. In any event, the same principles apply taking this as the unit of analysis as with taking the run as the unit. Assuming these new unit scores, summing over type, form a true Gaussian, judicious assignment of them to type could produce the effects uncovered at the series level.

Finally, it should be stressed that all these models are speculative, and at this stage there is no reliable basis for selecting among them. However, it is worthy of mention that a paranormal model need not assume a causal effect on the mechanism of the REG to account for the results of Jahn's experiments or, for that matter, REG data generally. The issue is important, because its resolution could influence the kinds of applications that might eventually spring from REG research.

NOTE

[1] The Z-scores are based on all trials, pooled over experimental conditions. When Z-scores were provided for only one condition, total Z-scores are estimated assuming chance scoring (Z=0) in the condition not reported. When no Z-score was reported at all, it is assumed to be zero.

Chapter 6

THE DELMORE EXPERIMENTS

Perhaps the most dramatic psi results to be produced by a single subject in the last twenty years have been provided by Bill Delmore (B.D.), who at the time was a law student at Yale University. He was tested in a series of formal restricted-choice ESP experiments at J.B. Rhine's Institute for Parapsychology in the early 1970s. The principal experimenters were Dr. B.K. Kanthamani, a psychologist and experienced parapsychologist, and Dr. E.F. Kelly, a cognitive psychologist who had only recently become involved in parapsychology. Secondary contributions were made by Dr. Irvin Child, a senior psychologist and professor at Yale. I will begin by describing the methods and results of the three main elements of the research program with B.D.

Single-Card Clairvoyance

Method

The single-card clairvoyance (SCC) method was devised to be the better controlled of the two card-guessing methods utilized (Kanthamani & Kelly, 1974a,b). Ten identical decks of standard playing cards were thoroughly mixed and scattered face down inside the bottom drawer of a desk. (These decks were periodically rescattered or replaced during the course of the experiment.) The experimenter was seated at the desk facing B.D., who was seated at the other side of the desk six to eight feet away. For each trial the experimenter "randomly" picked out a card from the pile in the drawer and, without looking at it, placed it inside a 3 3/4"x2 3/4" opaque folder. The experimenter then held the folder up to B.D. such that the back side of the card inside the folder was facing him. B.D. called out his response, after which the experimenter recorded the call, removed the card from the folder to observe it, and recorded it alongside the call on the record

sheet. For most trials, B.D. received immediate feedback as to whether the response was correct. In one series, B.D. was asked to make "confidence calls"; i.e., to note those trials where he felt particularly confident that the response was correct. (B.D. reported that this procedure was stressful, raising the possibility that he might have earmarked certain trials for extra effort and then chose them for confidence calls.) Each run consisted of 52 trials, with a break generally occurring after each 26 trials.

A preliminary experiment of 65 runs was conducted with Child as the experimenter, which later was reported as having been extended to 74 runs (Kelly, Kanthamani, Child, & Young, 1975). The main experiment consisted of 46 additional runs divided into four series. For Series 1 through 3, Kanthamani was the experimenter, whereas in Series 4 her husband, also a psychologist and parapsychologist, served as experimenter. In Series 2 through 4 the one who was not the experimenter was present in the room as an observer. Other observers were sometimes present as well. Confidence calls were invited in seven of the ten runs of Series 3. Feedback was withheld in 179 of the 2392 trials at B.D.'s request. The number of runs for each series was specified in advance.

The principal method of analysis was a procedure devised by R.A. Fisher (1924) for just this kind of task. Briefly, a composite Z-score is generated based upon the deviations from the expected values for each of the nine possible combinations of hits and misses on the attributes of number, color, and suit. The statistic was supplemented by various other chi-square tests based on the same general "goodness-of-fit" principle.

Results

The results of the Child experiment were marginal and not reported in great detail. The only significant finding was an excess of hits on suits, $Z=2.55$ ($p<.01$, one-tailed), over the 74 runs (Kelly et al., 1975). A sharp

upturn of scoring on the last nine runs of this experiment was also noted, although it is not stated whether the Fisher Z was independently significant for these runs. The upturn is probably attributable to the fact that these runs were administered during the same period of time as the more successful main experiment.

The Fisher Z for the main experiment (Kanthamani & Kelly, 1974a,b) was highly significant (Z=10.73) and was independently significant for each of the four series except the first. The effect was concentrated in an excess of "exact hits" (getting the card completely correct) which was three times the expected number (Z=13.00). The number of exact hits also exceeded that expected given the hit rate on the component attributes (Z=5.8). With the exact hits removed, there was still an excess of hits on numbers (Z>7) but, surprisingly, a significant deficit of hits on suits (Z=-3.2). B.D. scored somewhat better on the 179 nonfeedback trials than on a control set of 289 feedback trials from the same runs (no statistics reported), but interpretation of this finding is ambiguous because the nonfeedback trials were selected by B.D. at those times he felt "hot," as the authors note. Finally, of the 20 confidence calls, 14 were exact hits, which comprised over 50% of the 25 exact hits in the runs where confidence calls had been invited. The authors also note the related point that B.D.'s scoring success tended to occur in "bursts" throughout the experiment.

The data from the main experiment were later subjected to additional analyses in search of systematic errors in B.D.'s misses that might shed light on the cognitive processes involved (Kelly et al., 1975). To provide a baseline for these analyses, B.D. completed 75 runs in which the targets were slides of playing cards projected on the screen through a tachistoscope at 1/125 of a second. B.D. reported that the perception of these slides corresponded to the visual images of the targets he experienced during the ESP tests. Multidimensional scaling (MDS), canonical correlation, and other related techniques were applied to detect correspondences between the two

tasks in his pattern of errors in detecting the cards' numbers. Although the lower information rate prevented the demonstration of a statistically significant error structure in the ESP data, the pattern that did emerge was found to correlate significantly with the more reliable pattern uncovered in the visual data. Moreover, the correlation was found to be attributable almost entirely to the half of the ESP runs where the scoring level was highest, as one would expect. Also as one would expect, the confusions on both tasks consisted primarily of confusing the face cards with each other and the Ace, 2, and 3 with each other. MDS could not be applied to an analysis of errors regarding suits, but chi-square tests revealed in both sets of data a tendency for B.D. to confuse suits of the same color, the significant correlation again being attributable to the high-scoring ESP runs. (Further analyses by Kennedy [1979] revealed other confusion structures in the chance-scoring ESP runs [including, in some cases, confusing suits of opposite color], whereas no confusion structure seemed to be present in the low-scoring [psi-missing] runs.) Kelly et al. concluded that their results demonstrate "an overall structural resemblance between ESP errors and visual errors" (Kelly et al., 1975, p. 26) and they interpreted the finding as evidence that "on a significant fraction of occasions on which B.D. obtains ESP information, he encodes it in the form of visual imagery" (p. 27).

Following a discussion of possible artifacts (to be dealt with later), the authors concluded that "The procedures employed in these experiments seem sufficiently rigorous to create a strong presumption that the effects reported are genuine ESP effects" (Kanthamani & Kelly, 1974b, p. 24).

Shuffle Method

For each run, the experimenter shuffled one of over 24 decks of standard playing cards ("target" deck), to which B.D. was reported to have had no access, a minimum of ten times (Kanthamani & Kelly, 1975). B.D. then

shuffled another of these decks ("call" deck) as many times as he wished, attempting to duplicate the order of cards in the first deck. In the first two series, check-up occurred by the experimenter first recording the order of cards in the target deck and then the order of the call cards as B.D. turned them over successively. In the third series, the "calls" were not recorded or announced until B.D. had removed each card in the call deck from its pile and transferred it to a new pile face down. The effect of this exercise, suggested by B.D., was to delay somewhat his knowledge of the results. In Series 4, B.D. shuffled the cards inside a cardboard box with holes through which his arms could be inserted. The box was retained in Series 5 and 6, with B.D. also transferring the cards inside the box during check-up, and in Series 6 he was actually encouraged not to transfer the cards sequentially but to select a card from anywhere in the deck to match each target card announced by the experimenter. Confidence calls were also invited in Series 5 and 6. A few other modifications, one of which will be discussed later, were occasionally introduced.

The six series comprised a total of 55 runs of 52 trials each. Kanthamani served as the experimenter for all series, although various witnesses were said to be present during Series 4 through 6. The methods of statistical analysis were the same as described above for the SCC experiment.

Results

The Fisher Z for the total trials was highly significant (Z=12.88), and was significant for each of the six series separately. Even more so than with the SCC method, the significance was concentrated in an excess of exact hits amounting in this experiment to four times the expected number (Z=22). In contrast to the SCC series, the numbers of suit and number hits *per se* were close to chance expectation. Thus, the significant scoring would appear to be entirely attributable to exact hits. All 50 of the confidence

calls were correct and they comprised 68% of the 67 exact hits in the runs where confidence calls had been invited. Analysis of the misses by specialized chi-square methods revealed no evidence of systematic errors on either the number or suit attributes. The authors concluded that "The procedures of Series 1-4 appear to us to have been sufficiently rigorous to guarantee that the psi effects reported for them are genuine" (Kanthamani & Kelly, 1975, p. 216).

REG Experiments

The results of several preliminary series of experiments with B.D. were reported (Kelly & Kanthamani, 1972). Those using REGs are worthy of mention because of the relative security provided by this automated methodology.

Because these studies were preliminary, the methodological descriptions are rather sketchy. The most data were collected in ESP (precognition) using a four-button Schmidt REG (see Chapter 5) with a radioactive source of randomness. The authors stated that the device "in extensive tests covering millions of trials has never shown even minor departures from randomness" (p. 190), but details of these tests were not provided.

Several informal tests were recorded in which no hard copy of the results was obtained. The best controlled of these sessions, in which the tests were witnessed by J. B. Rhine and Helmut Schmidt, produced 180 hits over 508 trials (35.4%), with 25% expected by chance, which was highly significant (Z=5.4). The results of eight formal tests with automated recording of the results on paper tape yielded 1542 hits over 5377 trials (28.7%) with Z=6.24. Scores inclined over sessions from a nonsignificant 27.0% in the first session to 30.8% in the last session. The cumulative Z reached significance by the end of the second session.

The only other REG test involved B.D. and another subject jointly attempting to influence the output of a different Schmidt machine which produced binary trials at a rate of approximately 33 per second. A

1000-trial test yielded a modest but significant Z of 2.6.

Published Criticisms

The only scientist who has critically addressed the B.D. research program in print is Persi Diaconis (1978). In a paper published in Science, Diaconis did not discuss the experiments described above but instead focused on a demonstration of card guessing that B.D. had given earlier at Harvard in front of an audience. Diaconis was present at the performance and claimed that B.D. had used sleight of hand and the trick of "multiple end points" (not defining a successful outcome in advance) to create an illusion of psi. He then implied on the basis of these observations that the reports of the more formal experiments with B.D. cannot be trusted: "...the similarity of the descriptions of the controlled experiments with B.D....to the sessions I witnessed convinces me that all paranormal claims involving [B.D.] should be completely discounted" (p. 133).

In a rebuttal, Kelly (1979) argued that it was illegitimate to equate an informal, admittedly uncontrolled demonstration to formal, controlled experiments. He noted that the experiments were designed specifically to eliminate the kinds of artifacts that Diaconis claimed subverted the Harvard demonstration. For instance, multiple end points were excluded because the criterion for a successful outcome was specified in advance by the experimenters. As a secondary point, Kelly noted that Diaconis had not actually observed cheating but only inferred it.

In his reply to Kelly, Diaconis (1979) elaborated his position by stipulating that "ESP experiments done by known sleight-of-hand users must include, as part of the protocol, magicians skilled at detecting sleight of hand" (p. 30). In other words, the formal experiments with B.D. should be discounted because a skilled magician was not present to observe.

In his final rebuttal, Kelly (1980) argued that it is "...not all that difficult to design experimental conditions which are impervious to cheating

by any subject including a magician" (p. 123) and that significant results with a magician present could always be explained away by the argument that the subject was more skilled than the magician. Diaconis' (1980) final rejoinder introduced no new points on this issue.

Evaluation

I will begin with an evaluation of the research methodology based on the experimental reports and then consider the more far-reaching issues raised by Diaconis.

Sensory Cues

The SCC procedure as described in the report seems to successfully preclude sensory contact with the target card once it is placed in the folder. However, it is conceivable that, under certain circumstances, B.D. could have caught a brief glimpse of the card being transferred to the folder from the desk drawer. Specifically, if the subject were seated on the other side of the desk from the experimenter and had a pocket mirror in his lap, he might have been able to get a brief look in the mirror at the card being transferred. He would only need to do this on a few trials to obtain the reported results.

One feature of the data that is consistent with such a hypothesis is the similarity between the confusions structures on the ESP trials and tachistoscopic presentation of the targets. The kind of brief visual exposures of the targets in the tachistoscopic trials is very similar to the kind of brief exposures B.D. would have received were the sensory-cue hypothesis to be correct.

This hypothesis would be precluded, however, if it could be documented that the desk used for the SCC experiment had a back which extended down close to the floor. Fortunately, I was able to obtain some pertinent information on this point by interviewing Dr. Kanthamani, who is still at

the Institute for Parapsychology where the B.D. experiments were conducted. Without revealing my specific purpose, I asked her if she recalled and/or could show me the desk that was used for the SCC experiments. She stated that it was one of a set of very similar large, light brown wooden desks still at the Institute, but she could not recall which particular one it was. All the desks but one had backs. When I asked her specifically if the desk had a back, she said she was fairly certain that it did. She noted in particular that the desk was her own office desk and that she recalled having frequently rested her feet against the back of the desk when seated at it. Earlier in the interview she had also mentioned that B.D. would customarily sit facing sideways with his legs stretched out along the back of the desk. This would be a natural way to sit if the desk had a back, since if he were to sit facing the experimenter his legs would be jammed up against the back.

In summary, I came away from the interview with reasonable but not complete certainty that the desk used for the B.D. experiment did have a back and that my sensory-cue hypothesis was not applicable.

The shuffle procedure, on the other hand, seems somewhat more problematic, in that B.D. was allowed contact with the call deck after the target for each trial had been announced, thus allowing the possibility of either rearranging the deck or substituting new cards in order to fraudulently create hits. The authors acknowledge this as a problem in the last two series (in which B.D. had contact with the cards inside the box, outside the experimenters' view), in addition to the possibility of tactile heat cues from the cards, so-called "dermo-optic perception." However, the fact that no observers were present during Series 1-3 renders this hypothesis, while still not likely, more plausible than the authors acknowledge in their report. The best argument against this hypothesis is an appeal to two runs in Series 4 in which B.D. was not allowed any contact with the cards after the targets had been announced. The scores on these

runs were reported as being comparable to the other runs in the series, but it was not stated whether the scores of these runs were independently significant.

Sensory cues do not appear to be a problem in REG experiments with Schmidt machines, so barring unusual circumstances this criticism would not apply to this series.

Randomization

In the main SCC series, the authors submitted the actual sequence of targets in the 46 runs to analysis for singlet and doublet biases. A significant but modest singlet bias was in fact uncovered (statistical test not reported), which could easily happen if target cards from previous trials were not replaced in the pile in an entirely random manner. However, subsequent analyses revealed that these biases were not correlated with B.D.'s hits and cannot account for the results.

Biases due to inadequate randomization seem more plausible in the case of the psychic shuffle series. Although ten shuffles seem adequate in principle, in practice its adequacy rests on how the shuffles were performed. The problem is particularly acute in cases when the decks are ordered in corresponding ways to begin with, such as would be the case when decks are new. Unfortunately, it is not clear from the report how often such correspondences might have been obtained, nor were the actual sequences submitted to the kinds of analyses reported for the SCC series. The fact that the hitting was restricted to exact hits exclusively does seem consistent with such an interpretation.

The faulty-randomization hypothesis seems unlikely in the case of the REG series, but it would be desirable to have more information about how comparable the conditions in the randomness test were to those in the actual experiment, especially regarding rate of target generation.

Recording

Duplicate recording and counting of hits was not applied in any of the card-guessing series. However, in order to achieve the high levels of scoring obtained, errors of this type would have had to amount to gross negligence. Recording errors were apparently ruled out in the REG experiments using the paper tape. Errors in the counting of hits were also precluded, assuming that the Schmidt machine displayed hit totals.

Data Selection

Optional stopping was ruled out in all the card-guessing experiments because the number of trials per series was specified in advance. The authors appear to have been conscientious in reporting all the experiments conducted with B.D., including the exploratory experiments. In any event, the results were so highly significant that they could not easily be "washed out" by unreported negative findings.

Statistical Methods

The analyses of the data show a great deal of sophistication. The methods were standard, simple (except in the case of the secondary analyses of "confusions"), and appropriate. Biases due to uncorrected multiple analyses can be ruled out first by the extreme levels of significance obtained and second by the fact that the different methods of analysis yielded comparable conclusions.

Diaconis Critique

I agree with Diaconis to the extent that he argued there is prima facie cause for suspicion of subject fraud in the B.D. experiments. Although Kelly is correct that Diaconis inferred cheating by B.D. in the Harvard demonstration rather than directly observed it, I consider Diaconis' inference to be reasonable if not compelling. Especially in his first

example regarding the use of "multiple evidence," Diaconis cited not only a successful outcome but a whole pattern of behavior on B.D.'s part that suggests the use of procedures and principles that are very common in stage magic. Whether or not B.D. used tricks at Harvard, his behavior there indeed compelled particular circumspection during the formal experiments. At a minimum, the authors should have consulted with someone having expertise in conjuring about the adequacy of the experimental procedures. Although Kelly argued that it is not difficult to design experiments impervious to cheating by magicians, it is precisely Kelly's qualification to make that remark which Diaconis questioned. Even though the authors did consider and control for some possible forms of cheating, the absence in the reports of information that would render the mirror hypothesis inapplicable suggests that the authors were not sensitive to this particular possibility. This reinforces Diaconis' point, whether or not the mirror hypothesis is applicable in fact.

A second ground for suspicion is that in all the card experiments procedural modifications were instituted at B.D.'s request. The modification most likely to have impact on the results was the provision for B.D. to handle the call deck post-feedback in the psychic shuffle experiment. However, as noted previously, this modification was not in force throughout that experiment, and no one has yet suggested how B.D. could have used the modification to effect the results.

On the other hand, Diaconis should be faulted for apparently jumping to the conclusion that the Delmore experiments must be nonevidential simply on the basis of the Harvard demonstrations and his general impression of <u>other</u> parapsychologists. Such glib generalizations are clearly unwarranted. For instance, the level of performance exhibited in the formal experiment, while impressive in that context, would not be at all impressive in the context of a short public demonstration. Thus, if B.D. did use sleight of hand at Harvard, it may have been because he felt he needed to in order to achieve

the necessary outcome, whether or not he possessed, and knew he possessed, genuine psychic ability. However, the more important point is that the kinds of standard tricks Diaconis claimed were used at Harvard were precluded in the formal experiments, and Diaconis offered no counterhypotheses of his own to account for the results in these experiments.

In conclusion, whereas the authors should have exhibited more concern about the apparent magical skills of B.D. and less confidence in their own abilities to detect their use, Diaconis' critique lacks scientific weight. In the absence of even a plausible hypothesis as to how B.D. might have achieved his results fraudulently, they remain a genuine anomaly.

Chapter 7

CORRELATIONAL STUDIES

The research projects we have considered so far were designed primarily to demonstrate the existence of psi by producing statistically anomalous results under conditions that preclude orthodox explanations of those results. However, much of the research in parapsychology is conducted with the more modest objective of determining psychological correlates of psi scores or of determining how such scores are affected by the manipulation of experimental conditions. Such research does not tell us anything directly about the likelihood that the anomalies are paranormal, because the correlations uncovered could conceivably be predicted from orthodox as well as from paranormal theories. However, demonstrations of reliable relationships between psi scores and external variables are important for at least three reasons:

(1) They reveal at least a rudimentary coherence and lawfulness of the anomalies. When anomalies collected under diverse circumstances relate in the same way to external variables it suggests that the mechanism which underlies them is uniform; i.e., it is reasonable to talk about a coherent class of events;

(2) They may point to factors that if controlled or exploited could improve the reliability of psi scores;

(3) They can serve as the building blocks for theories about the anomalies or about how they interact with other psychological or physical processes.

It is important to stress that one need not have established "the existence of psi" (i.e., paranormality) for such research to be fruitful. Quite to the contrary, embedding the anomalies in a nomological net of functional relationships is one of the best ways of creating the basis for an incisive answer to the question of paranormality.

Progress in uncovering correlates of the anomalies in laboratory contexts has been excruciatingly slow. There are at least two reasons for this. The first is the lack of an adequate theory to guide the selection of variables. The second, and perhaps the more important, is the low internal reliability of psi scores, which is a by-product (at least in part) of the low signal-to-noise ratio. For example, even in Schmidt's ESP experiments with the REG, which were raving successes by parapsychological standards, the rate of successful guessing was only 1.5% above MCE. While this problem can be alleviated somewhat by collection of a large number of trials, this strategy can strain resources (especially in ESP experiments where individual trials cannot be accumulated rapidly) and it increases the difficulty of maintaining uniform control of extraneous biasing factors. To make matters worse, the reliability and validity of the psychological measure one seeks to correlate with psi scores is often far from ideal. This is not to suggest that the task is hopeless, but these factors may help account for the slow progress to date.

The above considerations suggest that correlations between psi scores and other variables are unlikely to be consistently replicable even if "real." In fact, most failures to replicate such effects can be attributed to error variance alone. This is not to suggest that unexpected correlations should be accepted at face value but rather that they should not be rejected out of hand. At the present state of parapsychology's development, the only way to reach a conclusion is to perform meta-analyses on large groups of studies addressing the same relationship to see if the distribution of outcomes departs from that expected by the null hypothesis. As we shall see later, this approach is not without problems of its own, but it is still "the best game in town" and in my opinion has provided useful hints about some future lines of investigation that might prove profitable.

Only a handful of external variables have been used in enough experiments to make meta-analysis feasible. There are but four variables

for which such analyses have been undertaken in any systematic fashion and these will be reviewed below. All are restricted to ESP scores as the dependent variable; only recently has an interest developed in uncovering the correlates of PK. All the predictors are psychological as opposed to physical variables. In all cases, the relationships have been classified simply as significant or nonsignificant; in no cases have attempts been made to assess the actual magnitude of the relationship or the "effect size."

Personality Correlates

Personality variables or "traits" can be defined as "behavioral dispositions or tendencies that are relatively stable over time for a particular individual and are so structured that each individual can be placed on a continuum for which that trait is an appropriate label" (Palmer, 1977, p. 175). A great deal of research has been done attempting to identify the underlying structure of personality. Factor-analytic approaches have tended to support the existence of two fundamental dimensions of personality, namely (1) "extraversion" and (2) "neuroticism" or "anxiety" (e.g., Cattell, 1965; Eysenck, 1960). It therefore is not surprising that these are the two traits which have been studied frequently enough in parapsychology to merit meta-analytic treatment.

Extraversion

From 1945 to the present there have been numerous attempts to correlate scores on ESP tests with scores on various personality tests claiming to measure extraversion. The most commonly used of these scales have been: the Cattell 16PF (and the version for adolescents called the High School Personality Questionnaire), the Maudsley (later, Eysenck) Personality Inventory (EPI), the Bernreuter Personality Inventory, the Guilford personality scales, and the Minnesota Multiphasic Personality Inventory

(MMPI).

The first meta-analysis of the extraversion-ESP relationship was by Palmer (1977), who sampled studies published in the major parapsychological journals. Using the experimental series as the unit of analysis, Palmer uncovered 33 series published in 11 reports which provided sufficient information to be evaluated. Palmer was not interested in the number of significant relationships per se, but rather the ratio of positive to negative relationships among the significant series as well as among all series. He found that 23 of the 33 series (70%) were in the positive direction (i.e., extraverts scoring higher than introverts) whereas all eight of the significant relationships were positive. This pattern proved to differ significantly from the pattern expected by chance, i.e., an equal number of experiments (and significant experiments) in the two directions. Palmer thus concluded that "there is evidence for a positive relationship between extraversion and ESP scoring" (Palmer, 1977, p. 186).

A more up-to-date meta-analysis was later reported by Sargent (1981). His survey included twelve reports not published at the time Palmer wrote his review, seven of which were from his own laboratory. He also included eight earlier reports not evaluated by Palmer. In seven of these cases, Palmer had not included the report because it gave no indication of the direction of the relationship. For reasons that are not clear, Sargent did not cite four of the reports cited by Palmer. In any event, the samples in the two surveys do not overlap as much as one might expect.

Unlike Palmer, Sargent was primarily interested in the proportion of significant outcomes. He also based his analysis on the number of reported relationships rather than the number of series, although these tended to be equivalent. From a total of approximately 54 relationships (it is not clear that this figure is exact), Sargent found 19 that were significant (35%) and 18 of these (95%) were in the positive direction.[1] This led Sargent to conclude "...that a real effect exists; extraversion is positively

correlated with successful performance in ESP tasks" (Sargent, 1981, p. 141).

Neuroticism

Neuroticism can be defined for present purposes as "tendencies toward maladaptive behavior caused either by anxiety or defense mechanisms against anxiety" (Palmer, 1977, p. 178). This definition subsumes anxiety as a special case of neuroticism, although the two terms tend to be used interchangeably in the parapsychological literature. All of the major personality scales cited above in the section on extraversion have subscales measuring neuroticism or anxiety, and scores on the subscales have also been correlated with ESP scores. Scales uniquely measuring neuroticism or anxiety that were used in more than one study were the Taylor Manifest Anxiety Scale and the projective Defense Mechanism Test (DMT).

Palmer's (1977) meta-analysis cited 21 reports that gave sufficient information for evaluation. When all series were considered, there was no evidence of a significant relationship between neuroticism and ESP scores. However, a post-hoc analysis revealed that a relationship did exist if the analysis is restricted to series in which subjects were tested individually or in pairs. (Palmer speculated that group testing might have alleviated state anxiety in the test situation among high-anxious subjects, thereby rendering trait anxiety an ineffective predictor.) Be that as it may, 18 of 24 series (75%) with subjects tested individually or in pairs yielded a negative relationship between neuroticism and ESP (higher scoring among less neurotic subjects) and all seven of the significant relationships were negative. These patterns differ significantly from the null hypothesis of equality, leading Palmer (1977) to conclude that "...there is evidence for a consistent negative relationship when Ss are not tested in groups" (p. 183) between neuroticism and scoring on ESP tests.

There also appeared to be differences in Palmer's survey in the "success" rates of the various predictor scales. The Manifest Anxiety Scale was the least successful and if anything tended to correlate positively with ESP scores. The most successful predictors were Cattell's neuroticism subscales and the Defense Mechanism Test. The latter is a projective technique in which the subject is asked to describe a threatening scene repeatedly displayed tachistoscopically at increasingly slower speeds. The defensiveness score is determined by how many exposures it takes for the subject to recognize the threat and the nature of the perceptual or interpretational errors made on preceding exposures. In a recent review of research correlating DMT scores with scores on restricted-choice ESP tests, which included seven experiments not reported at the time of Palmer's review, it was claimed that all ten experiments in the sample yielded a positive relationship; i.e., high ESP scores correlated with low defensiveness. In three of these studies the relationship was significant by a two-tailed test and in seven by a one-tailed test (Johnson & Haraldsson, 1984). The authors concluded modestly that "...the DMT seems to be a useful instrument in predicting the scoring direction in an ESP test" (p. 197).

Attitudes

The only attitudinal variable that has been extensively explored in relation to ESP scoring is belief in ESP, the so-called "sheep-goat" variable--i.e., "sheep" are "believers" and "goats" are "skeptics." Actually, the sheep-goat variable comprises four related attitudinal dichotomies which can be described in relation to orthogonal dimensions of generality and personal reference (Palmer, 1971): general-impersonal ("Do you believe ESP exists?"); general-personal ("Do you believe you have psychic ability?" or "Have you had psychic experiences?"); specific-impersonal ("Do you think the experiment you are now in can elicit

ESP?"); and specific-personal ("How well do you think you scored [will score] on this ESP test?"). One or more items of this type are incorporated into homemade rating scales. In most experiments the items are scored separately, and in no experiment published in the parapsychological literature has a scale been used which has undergone systematic test construction.

The one meta-analysis of the attitude variable was conducted over ten years ago by Palmer (1971). He used as his basis an experiment by Schmeidler (Schmeidler & McConnell, 1973/1958) comparing restricted-choice clairvoyance scores to attitudinal ratings on an item of the specific-impersonal type. The experiment consisted of seven series of individual testing and 14 series of classroom testing. Overall, 1308 subjects took part. Because undecided subjects were included among the sheep, only 505 subjects (39%) were classified as goats. Results were in the predicted direction in 18 of the 21 series (sheep scoring higher than goats) and results for all individual series combined and all group series combined yielded highly significant sheep-goat differences in each case. Palmer (1971) proceeded to review 22 experimental reports, including the original Schmeidler and McConnell report, which tested for attitude-ESP relationships. These were broken down into 24 "experiments" according to criteria that seemed reasonable from the structure of the reports but sometimes comprised more than one series. Formal meta-analysis was restricted to 17 experiments which could be uniformly reanalyzed by the Z-test of the number of hits per condition and used the standard card-guessing procedure with an expected mean score of five hits per run. Criteria were defined for classifying undecided subjects with respect to each of the four item types and applied as uniformly as possible throughout the sample. In cases where more than one predictor was employed, the direction was determined by majority vote; e.g., if two of three relationships were positive, the relationship was considered positive for

the study as a whole. In none of the experiments with multiple predictors were inconsistencies regarding significance of the relationship noted.

It was found that 13 of the 17 experiments (76%) were in the expected direction (sheep > goats). All six of the significant experiments were in the predicted direction. Moreover, it was shown that this distribution of outcomes closely approximated what one would expect if the true mean difference approximated the mean difference of +.17 hits per run found in Schmeidler and McConnell's combined group series, by far the largest sample available. Palmer (1971) concluded that "...the data presently available support the hypothesis of a genuine SGE [sheep-goat effect], although the relationship is very slight and difficult to demonstrate with small samples" (p. 405). Finally, a comparable rate of information was found among the experiments not included in the formal experiments (Palmer, 1971) and among experiments published in the early 1970s (Palmer, 1977).

Hypnosis

Although a great many variables have been incorporated in experimental manipulations designed to influence scoring in psi tasks, only one variable has been systematically manipulated in enough studies to be used for meta-analysis. What I mean by "systematically manipulated" is the results of an experimental treatment being compared to results in a control condition, either "within subjects" or "between subjects." The variable in question is hypnosis or, more precisely, hypnotic induction.

The attempt to facilitate ESP by means of hypnotic induction has a long history in psychical research (see Dingwall, 1968). However, the early research was poorly controlled and, because much of it was linked to the cult of Mesmerism, it was also tainted. J.B. Rhine did not find hypnosis helpful in facilitating card guessing and discouraged its use. However, beginning in the late 1950s a number of card-guessing studies employing hypnosis and reporting positive results began to appear in the literature.

Several studies also appeared using free-response methodology in the context of the "hypnotic dream," but few of these studies involved control conditions.

The experiments used fairly simple hypnotic induction procedures which were usually combined with suggestions for high scores on the ESP test. With one exception (Ryzl, 1962), deep hypnosis or training in hypnosis over a series of sessions were not employed. Subjects were ordinary volunteers claiming no special psychic or mediumistic talents and in most cases they were not even prescreened for hypnotic susceptibility.

The first meta-analyses of the hypnosis-ESP literature appeared in the late 1960s (Honorton & Krippner, 1969; Van de Castle, 1969). Each concluded that hypnosis indeed facilitated ESP scoring. However, the present review will focus on a more recent meta-analysis by Schechter (1984).

As was the case in the previous meta-analyses reviewed in this chapter, Schechter based his review on experiments published in the major parapsychological journals. He cited 20 reports which were classified as comprising 25 independent experimental series. Twenty of these were considered appropriate for the analysis, i.e., the experiment was designed to compare performance in the hypnosis and control condition, higher scoring in the hypnosis condition was expected, and the results were reported in such a way that the direction and significance of the difference could be determined. Schechter found that 16 of the 20 series yielded results in the expected direction (hypnosis > control) and that all seven of the significant outcomes were in this direction. Noting that the probability of seven of the 20 studies yielding significantly positive results by chance was slight (p=.000034 by an exact probability test), Schechter (1984) concluded that "...the apparent difference between ESP hitting after hypnotic induction and under control conditions is not likely to be a chance effect" (p. 6).

Criticism and Evaluation

There are three questions that I will try to address in this section. First, have the meta-analyses reviewed in the previous section demonstrated, within the samples evaluated, genuine relationships between ESP scores and the independent variables considered? Second, what, if anything, can be said about the interpretation or meaning of these relationships? Third, to what extent, if any, can the results of the meta-analyses be generalized beyond the samples included? In other words, is it reasonable to believe that the relationship will hold up in new samples?

Validity

It was stated at the beginning of the chapter that the objective of this review was not to confirm the anomalous nature of the ESP scores in the studies considered but rather to assess the reliability of the scores insofar as this follows from their consistent relationship with external predictors. For this reason, no attempt will be made to evaluate the possibility of artifacts in individual studies, an effort which would in any event be prohibitive from a practical standpoint because of the large number of studies involved. Moreover, several of these experiments have already been competently critiqued in a recent review by Akers (1984). Akers focused on the kinds of flaws or alleged flaws discussed in previous chapters of this review and found that most of the studies he considered were guilty of one or more of them. I think it is fair to say on the basis of his review that the prevalence and seriousness of the flaws found in the studies included in the present chapter closely approximate those of the ganzfeld experiments reviewed by Hyman (see Chapter 4). Thus, my analysis of the likelihood of the flaws uncovered by Hyman being the true explanations of the effects in the ganzfeld research applies to this chapter as well. Schechter (1984) included a discussion of these kinds of flaws in

his meta-analytic review of the hypnosis literature and found that none of them correlated significantly with the ESP scores.

Psychological Test Administration Following ESP Feedback. There is one flaw uncovered by Akers that deserves special treatment, however, because it affects the validity of the relationship between ESP scores and the predictor variables rather than the "validity" of the scores themselves. With reference to those experiments in which ESP scores were related to scores on personality or attitude questionnaires, Akers noted several instances in which the predictor scales were administered to the subjects after they had received feedback of their scores on the ESP test. This raises the possibility that subjects' psychological responses to the feedback may have influenced their responses to the items on the personality or attitude scale, thereby creating an artifactual correlation between the scores on those scales and the scores on the ESP test.

In attempting to assess the impact of this artifact on the studies which contributed to the previously reviewed meta-analyses, I first discovered that the order in which the personality and ESP tests were administered was often not reported, particularly in the studies conducted prior to 1970. Nonetheless, it still proved possible to come to conclusions about the possible impact of the artifact on the samples generally.

The artifact hypothesis can most clearly be rejected in the case of the sheep-goat effect. In only one of the studies included in the Palmer (1971) review could the Akers criticism apply (Nash & Nash, 1958). This was a nonsignificant study with results in the positive direction (sheep > goats) which was not, however, included among the 17 standard card-guessing experiments. All other experiments gave subjects no feedback of ESP scores before administering the attitude scale.

Of 49 distinct series culled from Palmer's and Sargent's reviews of the extraversion-ESP relationship, Akers' criticism clearly applied to 14 of

them (29%), whereas the description of methodology was insufficient to render an interpretation in another 16 (33%). However, the artifact proved to be unrelated to study outcome either with respect to significance or direction. The artifact applied definitely to only three of the 18 significantly positive experiments (17%) and definitely did not apply to 12 of them (67%).

The relationship for which the artifact appears most viable as an explanation is the one involving neuroticism. Of the seven significant confirmations in Palmer's (1977) review, the artifact is clearly applicable to three of them and may have been applicable to two others. In only three studies from the entire sample was the artifact clearly nonapplicable, and two of these provided nonsignificant reversals of the predicted trend (i.e., neurotic > nonneurotic).

There are several factors that militate against the Akers artifact accounting for the relationship, however. In each of the four significant and flawed studies involving objectively scored personality tests as predictors, subjects did not complete the personality test immediately after the ESP test but either at a separate session or in one case (Nicol & Humphrey, 1953) at home. Thus, any subtle mood shifts created by feedback of the ESP scores would have had time to dissipate. In the study cited by Akers to illustrate the potential effect of ESP feedback on psychological test scores, the psychological test (in this case a test of imagery skill, not personality) was given immediately after the ESP test (Palmer & Lieberman, 1975), and the test is especially susceptible to response biases. Three of the seven significant experiments were components of a series of four experiments by Kanthamani and Rao (1973). In one of these four experiments the artifact did not apply (the personality test was given before the ESP test), yet the neuroticism effect was still significantly confirmed. Since the procedure in the four experiments was otherwise quite similar, this result suggests that order of testing was not a crucial

variable.

The other two significant studies from Palmer's (1977) review involved the DMT. In neither study was it clear from the report whether the artifact was applicable or not. However, in the five studies from the larger, more recent DMT sample where the criticism clearly does not apply, all five were in the predicted direction and three of these were significant. Thus, again it would appear that the effect is not dependent upon the potential for the artifact being present. Particularly in the absence of any positive evidence that these personality scales are susceptible to influence by ESP feedback, it seems reasonable to conclude that the neuroticism-ESP relationship being attributable to the artifact suggested by Akers is unlikely. Nonetheless, he should be commended for bringing the potential artifact to our attention.

One other point about the methodology in the reviews requires brief comment. In all the reviews except Sargent's and Schechter's, a uniform criterion of significance was applied to all the studies considered. In Palmer's (1971) sheep-goat review this was the Z-test, because that was the only suitable alternative. In Palmer's (1977) neuroticism and extraversion reviews, conclusions were reached by averaging the results of the alternative analyses. There were no instances in which a study was classified as significant in which one analysis was significant and another analysis of the same relationship was not.

Interpretations

The existence of a correlation between ESP scores and a predictor variable says nothing directly about what psychological process or processes might be mediating it. Seeking first to establish the reliability of the correlations, parapsychologists have done little theoretically-oriented research designed to explain them. However, some limited speculations have

been published regarding the meaning of these relationships, all of which are based on the implicit assumption that ESP scores reflect paranormal processes.

Extraversion. Regarding the extraversion-ESP relationship, for example, two competing hypotheses have been proposed. The first, initially proposed by and based upon the psychobiological theory of Eysenck (1967), is that extraverts obtain higher scores on ESP tests because of a tendency toward low cortical arousal. A difficulty with this hypothesis is that it cannot explain why introverts tend to score below chance rather than at chance; at any rate, the positive deviation of extraverts seems no greater on the average than the negative deviation of introverts.

The second hypothesis is that extraverts are more at ease in the ESP test situation than are introverts. This hypothesis might also explain why less neurotic subjects and subjects who believe in the existence of psi also seem to achieve relatively high scores on ESP tests. In fact, there are some indications of overlap among these three predictors. One of the more successful predictors of ESP scoring has been the Cattell scales, on which the extraversion and neuroticism subscales are correlated and contain some overlapping items. This suggests that the low-scoring ESP subjects in these experiments may have been the introverted subjects who also showed signs of neuroticism. Such subjects would also be expected on commonsense grounds to be the most uncomfortable in a psi test situation, but no direct evidence of this has been provided. Thalbourne (1981) has consistently found a low positive correlation between extraversion and belief in ESP, suggesting some overlap between these variables as well.

Hypnosis. Alternative hypotheses also exist about the facilitative effect of hypnosis on ESP scores. One hypothesis is that hypnosis facilitates psi by putting subjects in a state of relaxation in which

attention is focused on internal processes, whereas the other attributes psi facilitation to the implicit or explicit suggestions of scoring success associated with the hypnotic inductions. Only one study compared these elements directly in a factorial design (Casler, 1962), and a reanalysis of this experiment which I conducted using analysis of variance supported the suggestion interpretation. However, trends in some of the other studies seemed to favor the state interpretation (Honorton & Krippner, 1969). The suggestion interpretation implies a possible link between the hypnosis and sheep-goat effects, in that hypnosis can be seen as a manipulation of the same belief variable that is simply being measured in the sheep-goat studies. A problem with this analogy is that the hypnotic suggestions have primarily manipulated belief in one's own ability to achieve a high score, whereas items which ask this question directly have been relatively poor predictors in the sheep-goat experiments. However, the failure of this item to discriminate scoring in sheep-goat experiments may be attributable to the highly restrictive range of responses to this item in most such experiments; ratings of high confidence are quite rare. Attempts to manipulate belief in ESP by means other than hypnosis have yielded mixed results (Layton & Turnbull, 1975; Taddonio, 1975).

Somewhat more arcane alternative hypotheses having to do with inadequacies in the design of many of the hypnosis-ESP experiments have recently been discussed by Stanford (in press). He noted first that in the great majority of the studies within-subjects designs were employed. He argued that subjects might be expected to score better in the hypnosis condition than in the control condition simply because of demand characteristics, i.e., the subjects knew they were <u>supposed</u> to score better under hypnosis and adjusted their expectations and motivation to perform accordingly. He also noted that in seven of these studies, there was either incomplete or no counterbalancing, with the control condition tending to come first in all cases. The fact that these studies tended to be less

successful than those which employed proper counterbalancing suggested to Stanford the possibility that hypnosis might only be successful when the hypnosis trials are presented first. Moreover, all five studies which avoided such problems by using between-subjects designs failed to assign subjects randomly to the hypnosis and control conditions, raising the possibility that differences in subject characteristics might be the real cause of the effects observed. Finally, in all studies the experimenter administering the ESP test was not blind to the condition assigned to the subjects they were testing, which raises the possibility that the experimenters may have unwittingly provided more encouragement to subjects in the hypnosis condition or otherwise interacted with them differently from subjects in the control condition.

Stanford's point about the problems associated with within-subjects designs is well taken, especially in view of the fact that when subjects are asked to perform in two psychologically distinct conditions in other kinds of ESP experiments they tend to score above chance in one condition and below chance in the other (Rao, 1965). Parapsychologists have traditionally felt that differences in motivation can affect psi performance (e.g., Rhine, 1948), although the actual empirical evidence for this proposition is scant (Weiner & Geller, 1984). Nonetheless, the possibility that demand characteristics account for much if not all of the hypnosis-ESP effect must be taken seriously.

The seriousness of the lack of counterbalancing in some of the studies is substantially ameliorated by the failure of these studies to achieve the same rate of success as those more properly counterbalanced. Even if Stanford's suspicion that the hypnosis-ESP effect is limited to cases where the hypnosis runs were not preceded by control runs is correct, the basic integrity of the effect would not be challenged. Most process-oriented research in parapsychology derives its conclusions totally or at least in part from the first (and only) ESP test the subject undertakes in a session,

and there is almost never a firm basis for inferring that a relationship would hold up over repeated testing, especially given the low reliability of ESP scores. There is some evidence for decline in ESP scores during the course of a session (Palmer, 1978) that might also lead one to doubt whether scores obtained later in a session would correlate as well with predictors as scores obtained earlier in a session.

The seriousness of the lack of random assignment of subjects to conditions in the between-subjects experiments depends upon the nature of the alternate procedure employed. These vary from subjects assigning themselves to conditions at one extreme (Moss, Paulson, Chang, & Levitt, 1970) to the experimenter assigning subjects alternately to conditions at the other (Casler, 1962), the latter being a perfectly acceptable procedure in this reviewer's judgment. Unfortunately, the one between-subjects study which provided a significant superiority of the hypnosis over the control condition (Sargent, 1978), seemed (as far as one can tell from the report) to use one of the more arbitrary and therefore suspect assignment procedures. On the other hand, the similar distribution of outcomes of the between- and within-subjects experiments suggests a common mechanism in both, and this would militate against subject-population differences being the effective cause. However, it certainly cannot be ruled out.

Finally, given the extensive evidence for the experimenter expectancy effect in psychology (Rosenthal, 1966), one cannot discount the possibility of the hypnosis-ESP effect being somehow related to the experimenters not being blind to the experimental condition. However, it should be noted that it is difficult (although probably not impossible) to guarantee such a blind in an experiment where one condition involves the subject being in an altered state likely to be identifiable by the person administering the test. It also should be noted that insofar as these demand characteristics affect the expectations and motivations of the subject, they cannot be clearly distinguished from one of the direct functions of the hypnotic

induction, which is also to increase the subject's motivation and expectation of success. This was previously discussed as one of the "nonartifactual" interpretations of the hypnosis-ESP effect (p. 152).

Belief. The sheep-goat effect has generally been interpreted as the ESP scores reflecting the needs and motivations of the respective subgroups; i.e., sheep and goats score in such a way as to confirm their previous belief systems (e.g., Palmer, 1972). Empirical evidence for this proposition has been provided by Lovitts (1981), who found a reversal of the sheep-goat effect among subjects who were led to believe that high scores would favor an alternative interpretation to ESP, namely subliminal perception. Although critics have not addressed the correlational literature from this point of view, one could reasonably hypothesize that sheep would be more likely than goats to be motivated to obtain high scores by cheating. However, this hypothesis would not account for the significant psi-missing by goats in some studies and in most studies it is not clear how subjects could have cheated.

In conclusion, interpretation of the correlational patterns discussed in this chapter must be considered speculative at this time. Furthermore, the viability of these or any other interpretation is obviously affected by the generality of the patterns themselves. It is to this question that we now turn.

Generality

If the findings of the meta-analyses discussed in this chapter are both valid and the samples on which they were based representative of the general population of such samples, then the trends should be preserved in experiments conducted after the meta-analyses were undertaken. One of course should not expect any particular study to significantly confirm the relationship, but the predicted trend should appear across a group of such

studies.

The ideal way to confirm the generality of the patterns would be to commission a planned series of replications from a variety of laboratories. Such a project presents obvious logistical problems and in any event has yet to be undertaken. Although informal surveys have seemed to confirm the continuation of the patterns found in some of the earlier meta-analyses (e.g., Palmer, 1977, 1978), these surveys are suggestive at best.

Some discouragement regarding the potential generality of the patterns is provided by a series of eight ESP experiments conducted by Michael Thalbourne and colleagues (Thalbourne, Beloff, & Delanoy, 1982; Thalbourne, Beloff, Delanoy, & Jungkuntz, 1983; Thalbourne & Jungkuntz, 1983). Subjects were mostly naive college students, high-school students, or volunteers from the community, with sample sizes ranging from 14 to 246 (mean=117.38). The dependent variable was the score on a ten-item restricted-choice clairvoyance test called "Consumer's Choice" in which the targets were brand names of consumer goods. The independent variables were belief in ESP and extraversion. Belief was measured by a ten-item sheep-goat scale created by the authors, and extraversion was measured by relevant subscales of either the EPI, 16PF, or MMPI. For purposes of analysis, subjects were divided into two groups on each of the independent variables. In the first two studies dichotomization was based on the means of the student populations from which the sample was derived; in the later studies the method of dichotomization was not reported but would seem to have been comparable to the original method.

Four of the eight sheep-goat relationships were in the predicted direction and one of these four was significant. This pattern seems tilted in the right direction (thanks to the one significant study) but is hardly a ringing confirmation of the sheep-goat effect. More distressing is the fact that in six of the eight studies, introverts scored higher than extraverts. None of these relationships was significant by the primary analysis, but the

last two experiments were each significant (favoring introverts) by a secondary, but more sensitive, correlational analysis. Although this pattern is not definitive even within Thalbourne's samples and is not sufficient by itself to overcome the large pattern in the opposite direction reviewed by Sargent, it does raise legitimate questions about the generality of this positive extraversion-ESP relationship.

It is now time to step back and examine the appropriateness of previous meta-analyses as bases for inferences to new samples. To put this somewhat differently but also more usefully, how can the populations which the meta-analytic samples truly represent be defined?

Only in the Palmer (1971) review were the publication sources sampled explicitly named. However, it is clear that the sources in all cases consisted almost exclusively of English-language parapsychology journals and abstracts of convention proceedings. The reviewers did not exhaustively consult nonparapsychological psychology journals and other scientific journals. The authors who publish in these journals are generally skeptical regarding psi and such journals tend to favor articles supporting the skeptical viewpoint. This failure to review exhaustively these journals may have led to an overestimate of the number of significant confirmations of the "expected" relationships, and the author knows of a couple such cases he "missed" in his reviews. However, the number of experimental parapsychology papers published in nonparapsychological journals is so small relative to the number published in parapsychological journals that the bias is slight.

The File-Drawer Problem. Thus, the sampled experiments are reasonably representative of the relevant published experiments conducted by those parapsychologists who were conducting experiments of those types prior to the date of the review. But what about experiments conducted but not published, the so-called "file-drawer problem"? This is a potentially greater problem in the attitude and personality studies than in the hypnosis

studies (and in the ganzfeld experiments discussed in Chapter 4) because the former type of experiments is easier and more economical to conduct. Sheep-goat questions, in particular, are economical to introduce and could easily go unreported in studies when they yielded no significant relationships.

It is well known that in the behavioral sciences "significant" experiments are more likely to be both submitted and accepted for publication than "nonsignificant" ones. However, this factor is not relevant to most meta-analyses reviewed in this chapter because they were concerned with the ratio of positive to negative relationships among the significant studies or among all the studies in the sample. Publication bias with respect to the direction of relationships is much less plausible than publication bias with respect to the significance of relationships, especially prior to the publication of the meta-analyses (which alerted investigators to the importance of directional trends). However, even allowing for the new awareness, it is difficult for me to conceive of a significant reversal of a relationship being suppressed, given the on-going mentality and practices of both researchers and the parapsychological journals. The major journals are forbidden by the Parapsychological Association from rejecting papers due to nonsignificant results and this would extend by implication to reversals as well. Actual data about unpublished experiments would obviously be superior to the preceding ruminations, but I think it is a good bet that the relationships uncovered in the meta-analyses are not attributable to the file-drawer problem.

The Experimenter Effect. There is another factor which I think is much more likely as compromising the generality of these patterns, and that is the so-called experimenter effect. It is widely agreed among both psi proponents and critics that some investigators are consistently more able than others to obtain significant results in their experiments. This factor

was not taken into account in any of the meta-analyses under consideration, since separate series by the same investigator or laboratory were treated as independent. There is nothing illegitimate about this, but it does obscure the possible mediating effect the identity of the investigator might play in accounting for the relationship. In all the reviews there were several instances where a particular investigator contributed more than one series. The ratio of investigators to series was: 15/24 in Palmer's (1971) sheep-goat review (12/17 for the standard card-guessing experiments), 7/24 in Palmer's (1977) neuroticism review (minus group experiments), 8/33 in Palmer's (1977) extraversion review, 19/53 (approx.) in Sargent's (1981) extraversion review, and 10/20 in Schechter's (1984) hypnosis review (scorable studies). Discounting Palmer's extraversion review, which is largely subsumed by Sargent's, the number of investigators is approximately 42% of the number of series, an average of 2.37 series per investigator.

The extent to which the significant relationships depended on a small number of investigators varies from review to review. This factor is revealed most clearly by considering the studies which significantly confirmed the general trend. The least effect of investigator uniformity was in Palmer's (1971) sheep-goat review, where six of the seven significant outcomes were obtained by different investigators. The greatest effect was in the neuroticism review of Palmer (1977), where only three of the seven significant studies were by different investigators and a single investigator (Kanthamani) was involved in five of them. The situation was not quite so severe in Palmer's (1977) extraversion review, where the eight significant studies were contributed by five investigators, but three were again contributed by Kanthamani. In Sargent's (1981) extraversion review, 11 of the 19 significant outcomes were by different investigators. However, six were by Sargent himself. Finally, four investigators produced the seven significantly confirmatory outcomes in Schechter's (1984) hypnosis review, three being produced by a single investigator.

It is especially noteworthy that the two investigators who contributed most heavily to the personality-ESP patterns both have excellent track records as "psi-conducive" experimenters in other contexts. Sargent is known for getting significantly overall positive results in ganzfeld experiments (although it should be noted that the significant personality-ESP correlations were generally obtained in these same experiments), and Kanthamani was the principal investigator in the Delmore experiments (see Chapter 6).

The final way to look at experimenter uniformity in these reviews is to calculate the proportion of experimenters who failed to obtain even a single significant confirmatory result. There were nine of fourteen (64%) in Palmer's (1971) sheep-goat review, four of seven (57%) in Palmer's (1977) neuroticism review, three of eight (38%) in Palmer's (1977) extraversion review, eight of 19 (42%) in Sargent's (1981) extraversion review, and six of ten (60%) in Schechter's (1984) hypnosis review. Discounting Palmer's extraversion review which is largely subsumed by Sargent's, 54% of the sampled experimenters failed to obtain a significant result.

Taken as a whole, this analysis suggests that the significant outcomes were not evenly distributed among the investigators responsible for them. This is the case in all the meta-analyses except possibly the sheep-goat one. To the extent this is true, it suggests that an important factor in determining the probability that any of these other effects will be replicated is the identity of the investigator attempting the replication. This point is relevant to the case of Thalbourne, whose unsuccessful series of replications was discussed earlier. Thalbourne's research, because of its recency, has not been included in any of the reviews under consideration and he is not known as a "psi-conducive experimenter" in other contexts.

Nonetheless, some encouragement can be derived from the fact that 46% of the experimenters sampled did achieve a significant result. However, it must be remembered that even considered as a whole, these experimenters are

a rather unique bunch. With but one known exception the experimenters were all parapsychologists positively inclined toward the existence of psi. Most have a keen interest in the subject. Many are in the field precisely because they have had at least a modicum of success in producing psi effects in the laboratory. In other words, they define a highly specialized population, the definition of which probably cannot be legitimately extended to include the garden variety psychologist who must someday succeed in replicating these patterns if they are ever to achieve the stature of genuine psychological laws.

There is not yet an adequate explanation of the "experimenter effect" in parapsychology. Speculative hypotheses include differences in experimenter honesty and competence, different social skills in handling subjects, use of subtlely different subject populations, and paranormal mediation by the experimenter (i.e., it is the experimenter, not the subjects, who produces the psychic effect). It is probably naive in any case to suspect a bivariate correlation between ESP scores and any external variable not to interact with other variables (Palmer, 1977). However, when one such variable is the investigator, special problems are created, since the important scientific principle that an effect (whether simple or complex) can in principle be replicated by any competent researcher is undercut. The message that should be drawn from this is not that psi is an artifact or that parapsychology is a waste of time, but rather that priority must be given to understanding the experimenter effect so its deleterious effects can be circumvented.

Chapter 8

PSI-MEDIATED INSTRUMENTAL RESPONSE

Critics of parapsychology often claim that the field lacks any serious theorization (e.g., Alcock, 1981). Although the level of theoretical development in parapsychology is indeed primitive in comparison to other sciences, this criticism is an overstatement. In order to illustrate the maximum level of theoretical development which has so far been attained in parapsychology, I have decided to review a research program undertaken in the 1970s by Dr. Rex Stanford related to a concept called psi-mediated instrumental response (PMIR). Stanford was trained as a psychologist and his research program is typical of what one finds in psychology. It addresses psychological issues pertaining to psi, in particular how psi is processed cognitively and how it interacts with the needs and motivations of the organism.

Stanford's research program contains a number of features normally associated with a sound theoretically-oriented approach in psychology. These include the following:

(1) Development of a model of broad scope that integrates previous experimental findings as well as anecdotal observations;

(2) Presentation of the model as a series of clearly stated propositions that are testable;

(3) Experimental testing of these propositions by means of an appropriate and standardized methodology.

As is perhaps evident from the discussion of other research projects in this review, this degree of logical development is not representative or typical of parapsychology in general. Although low-level theorizing and hypothesis testing is rather common in parapsychology, it does not possess the degree of systematization evident in Stanford's program.

The PMIR Model

The essence of the PMIR model is summarized as follows by Stanford (Stanford & Stio, 1976, p. 55): "...[an] individual, through extrasensory means, actively scans his environment for objects and events (or information related thereto) which are relevant to his needs and that when such information is discovered he tends to respond to it in accordance with his typical dispositions toward such objects and events."

What is novel about this proposition is the assumption that the individual is constantly and actively seeking out information in the environment by means of a paranormal process (Stanford, 1974a). In order that this scanning not interfere with normal cognitive activity, it is of course necessary to postulate further that the scanning is unconscious. Specifically, Stanford postulated that the scanning and the response made as a result of the need-relevant information obtained by the scanning (i.e., the psi-mediated instrumental response) can occur "(a) without a conscious effort to use psi; (b) without a conscious effort to fulfill the need... (c) without prior sensory knowledge...of the need-relevant circumstance; (d) without the development of conscious perceptions (e.g., mental images) or ideas concerning the need-relevant circumstance; and (e) without awareness that anything extraordinary is happening" (p. 45). These assumptions vastly broaden the population of potential psi events, which traditionally have been restricted to cases where a person has a conscious "psychic" experience (spontaneous experience) or consciously intends to use psi (as in an experiment). The PMIR model subsumes such cases but deals with others as well.

A typical "PMIR experience" cited by Stanford is that of a couple who wanted information about good vegetarian restaurants in Washington, D.C. While eating lunch at a restaurant en route to Washington, they chose to sit in a booth where they overheard a conversation between the people in the adjacent booth which provided the needed information. According to the

model, the couple was scanning the environment psychically for the needed information and, having received the information, responded in such a way as to fulfill the need (i.e., by choosing the "right" restaurant and the "right" booth at the "right" time), all of this being done without awareness that they were using psi to obtain this information.

Stanford also postulated certain psychological mechanisms which lead to psi-mediated responses. Following an earlier theory by Roll (1966), he hypothesized that psi does not introduce new cognitions into the individual as such but rather that it facilitates or triggers the selection of cognitions (e.g., memories) or behaviors which already exist in the individual's repertoire. Stanford proposed that this response is accomplished as economically as possible through a variety of mediating vehicles, including (1) modification of the timing of an already selected response; (2) forgetting or remembering to do something; (3) making a mistake (e.g., dialing a "wrong" number); (4) a thought coming to mind that leads by a normal chain of associations to the intention to make the response; and finally (5) the direct (conscious) cognition of the need-relevant circumstance, as in a traditional "psychic experience."

The strength of the disposition toward PMIR was postulated to be associated with "the importance or strength of the need(s)," the degree of relevance of the need-relevant object or event, and the "closeness in time of the potential encounter with the need-relevant object or event" (Stanford, 1974a, p. 45). The likelihood or effectiveness of PMIR was postulated to be influenced by certain situational and/or psychological factors, in particular competing cognitive activity that increases the rigidity or stereotypy of thought or behavior. Certain psychological traits such as neuroticism, guilt, or a low self-concept may cause the individual to use PMIR masochistically to counter his or her apparent best interests.

In a later paper, Stanford (1974b) extended the PMIR concept to cover PK. He postulated that the psi-mediated response could be psychokinetic

(i.e., paranormal) as well as normal in nature. Just as in the case with ESP, the model assumes that PK can function unconsciously; i.e., "PK...can occur as a response to extrasensory or sensory information which has never been in the conscious focus of the PK...agent..." (p. 350). This of course implies that PK can occur nonintentionally. Even when PK is used intentionally, the model postulates that it is facilitated "when the goal event is not in the conscious focus of the PK agent [although it] has definite motivational salience" and in particular when the goal event has "just left the focus of consciousness without having been realized" (p. 349). In fact, the probability of PK is actually "reduced during [a] period of focused attention and wishing" (p. 350). Shifting responsibility and capacity for PK away from oneself and onto an external agency (as, for example, one does in prayer) tends to discourage direct focusing of attention on the problem and is thus PK-facilitory.

Stanford also noted that those forms of telepathy in which the agent actively "sends" information to the percipient can be construed as a subcategory of PK. This so-called "active-agent telepathy" was renamed by Stanford "mental or behavioral influence of an agent" (MOBIA), which he postulated is "the most frequent PMIR function of PK" (p. 349). MOBIA follows the same laws within the model as do other forms of PK.

The above discussion represents a condensation of 18 formal postulates presented in the two papers, and the quotations I have cited were taken directly from those postulates. Although not stated explicitly in the papers, the model is obviously linked to the basic principles of reinforcement theory in psychology and thus provides at least a potential bridge between psychology and parapsychology.

In the course of his presentations Stanford cited numerous experiments in the parapsychological literature in which psi effects occurred in the absence of direct intention by the subject or were influenced by aspects of the target situation of which he or she was not aware. Examples include

effects on ESP scoring as a function of whether or not target cards were paired with erotic photographs (Carpenter, 1971), performance on a classroom exam being influenced by the unknown presentation of hidden answers to some of the questions (Johnson, 1973), and success on dice-throwing tasks when the subject was unaware which face had been chosen as the target (e.g., Fisk & West, 1958).

Methodology

Stanford developed a standardized methodology to test various propositions of the PMIR model. In general terms, the approach was to have subjects engage in a covert psi task, i.e., a task which the subject did not realize involved testing for psi. If the subject's performance on this task met a certain prespecified criterion, he would be allowed to escape or avoid some unpleasant, boring task and engage instead in a pleasant, interesting task. Thus, if the responses on the covert psi task were indeed psi mediated, they could determine an outcome relevant to the subject's needs.

Stanford has published five experiments using this methodology (Stanford & Associates, 1976; Stanford & Rust, 1977; Stanford & Stio, 1976; Stanford & Thompson, 1974; Stanford et al., 1975). Subject samples ranged in size from 29 to 72 and consisted exclusively of college student volunteers. In all but one study (Stanford & Rust, 1977) the subjects were exclusively males.

In all but one of the experiments the covert psi test involved ESP. In these cases the test was introduced to subjects as a standard test of word association. In this type of test the subject is presented with a taped stimulus word and is asked to respond with the first word that comes to mind. Thirteen words were used, the first three serving as buffers to acclimate the subject to the procedure. The remaining ten words were chosen so as to have a primary response (that is, the normatively most common response) of moderate strength. The parameter of interest was the response

latency--how long it took the subject to respond--which was recorded on a microswitch-activated electric timer. In most of the studies this timer was accurate to .1 sec although in the Stanford and Rust experiment the accuracy was improved to .01 sec.

For each subject, one of the ten words was randomly chosen by means of a random number table to be the key word. If the response latency on this word was the shortest of the ten (or, in some predesignated cases, the longest) or was tied for this distinction, the subject was subsequently invited to engage in a "pleasant" task. Otherwise, the subject was asked to engage in an "unpleasant" task. The subject was told nothing about these subsequent tasks at the time of the word-association test and was not told of any contingency involving the test.

In terms of the PMIR model, the specific cognitive mechanism available to the subject for determining his fate is what Stanford calls the "unconscious timing mechanism." In other words, it is the psi-mediated timing of the response, rather than its occurrence per se, that is instrumental.

In all but one experiment, the "pleasant" task consisted of male subjects' rating photographs of nude or semi-nude women. In the remaining case, the male subjects received relaxation suggestions from an attractive female research assistant (Stanford & Stio, 1976). The most common "unpleasant" task was for subjects to use a photocell stylus to track a small patch of light on a pursuit rotor turning at a boringly slow speed. Other unpleasant tasks were circling any of three designated letters the subject should find on three sheets of paper filled with all the letters of the alphabet in random order (Stanford & Thompson, 1974) and an ESP card-guessing task (Stanford & Stio, 1976)--an interesting commentary on how Stanford views such tests! In all cases the task was introduced as a genuine part of the experiment, designed to collect useful psychological data.

In one experiment (Stanford & Rust, 1977), the subject who experienced the pleasant or unpleasant task was not the same subject who had taken the word-association test. This experiment was designed to determine if PMIR might be used altruistically.

In the one PK experiment (Stanford et al., 1975), the subject began by engaging in the pursuit-rotor task. Unknown to the subject, an REG in the next room was left running. The machine was programmed to have $P=1/6$ for a hit and generated trials at a rate of one per second. The machine produced a maximum of 2700 trials per subject, equivalent to 45 minutes on the pursuit-rotor task. (This was done in five-minute intervals with one-minute breaks in between). The REG counted hits in blocks of ten. When and if there were seven or more hits in a block ($p<.0003$), the subject was removed from the pursuit-rotor task and allowed to engage in the picture-rating task. The chance probability of this occurring for any subject was .072.

Scoring

In the ESP test, the primary dependent variable was a standardized transform of the response latency on the key stimulus word. This was obtained by applying a log transform to all the response latencies and subtracting for each subject the latency to the key word from the mean of all ten latencies and dividing by their standard deviation. In the REG experiment, the dependent variable was simply the proportion of hits produced by the REG while it was active.

In all the experiments a possibly more appropriate, although less sensitive method, would have been the number of subjects who actually escaped the unpleasant task. Stanford did not evaluate this measure in the first three ESP experiments because the prevalence of ties in the response latencies of particular subjects meant that the chance probability of escaping the unpleasant task would vary from subject to subject. The more precise timing measure eliminated this problem in the Stanford and Rust

(1977) experiment, so the discrete dependent variable (number of subjects escaping the unpleasant condition) was used in this case, although it is not clear from the report whether this or the standard scores were construed as the primary measure. The discrete scores were also computed in the PK experiment (Stanford et al., 1975), but in this case the continuous (proportion of hits) scores were stipulated as primary. The standard scores were evaluated using common parametric tests such as *t* tests and analysis of variance. Exact probabilities were used to evaluate the discrete scores.

Results

In terms of overall scoring, the results of these experiments are not particularly impressive. Only in the PK experiment were the overall scores significant on the continuous measure. However, the discrete scores were significant in both studies where such scores were computed (Stanford et al., 1975; Stanford & Rust, 1977). In all five studies all the results reported were in the predicted direction--i.e., above chance. However, in most of the experiments the psi scores were related to independent variables measured for the purpose of testing specific propositions of the PMIR model. The results of these tests can be summarized as follows:

(1) *Overt Psi Tasks*. In two experiments, the proposition that unconscious and nonintentional psi is the same process as conscious and intentional psi was tested by correlating scores on the covert psi test to scores on a standard (overt) psi test conducted at the same session. In the Stanford and Thompson (1974) experiment the overt test was a precognition task in which the subject had to predict which segments of a printed "radar screen" would later be randomly chosen to contain targets. The correlation was positive and significant, confirming the hypothesis ($r=.39$, $p<.025$, one-tailed). In the PK experiment, subjects took an 80-trial overt PK test on the same REG used in the covert test. The correlation in this case was

in the positive direction but not significant (r=.20).

(2) Ready Responses. The PMIR model postulates that PMIR is more likely to be facilitated by readily available cognitions than by more submerged ones. In the context of word-association theory, this means that primary responses (the most common responses in the population according to published norms) to the stimulus words are more likely to be good mediators than are other responses. Since primary responses are generally associated with short reponse latencies, Stanford reasoned that PMIR would be more likely to occur in conjunction with short-latency responses (likely to be primary) than long-latency responses. In the Stanford and Stio (1976) experiment, this hypothesis was tested by manipulating whether the shortest or longest latency was instrumental in escaping the unpleasant task. As predicted, the mean standard score for the fast-contingency condition was significantly above chance ($p<.02$, one-tailed) and significantly higher than the mean standard score for the slow-contingency condition ($p<.02$, one-tailed).

(3) Need Strength. The PMIR model postulates that the disposition toward PMIR is positively related to the strength of the need served by it. Capitalizing on the erotic nature of the picture-rating task, Stanford and Stio (1976) atttempted to manipulate need stength (orthogonally to response-speed contingency) by having half of their subjects listen to an erotically arousing record before engaging in the word-association test, the idea being that the record would increase the "need" to participate in the picture-rating task. The other subjects heard the record after the word-association test. The manipulation failed to affect psi scores, but the authors suggested retrospectively that the record may not have really been erotically arousing.

In another experiment (Stanford & Associates, 1976), the authors attempted to manipulate need-strength by merely manipulating the sex (and therefore the sexual attractiveness) of the experimenters conducting the word-association test. As predicted, subjects tested by female experimenters scored higher than subjects tested by male experimenters ($p=.025$, one-tailed), although the mean score of the former subjects was not significant by itself. These subjects also scored significantly above chance ($p<.05$, one-tailed).

Although not construed as a test of the need-strength hypothesis, the results of the two experimenters in the PK experiment (Stanford et al., 1975) were also compared. Results collapsed over both the overt and covert PK tests were significantly higher for subjects tested by the more extraverted of the two experimenters ($p<.01$). Subjects tested by this experimenter also scored significantly above chance as a group on both the covert ($p<.01$) and overt ($p<.05$) tasks. Parapsychologists generally assume that extraverted experimenters are better able to motivate subjects in psi experiments than are introverted experimenters (e.g., Sargent, 1980).

(4) Self-Concept. The PMIR model postulates that a positive self-concept leads to use of PMIR in support of the subject's self-interest whereas a negative self-concept can lead to the reverse. Stanford and Associates (1976) attempted to create a positive self-concept in half of their subjects (orthogonal to the need-strength manipulation) by giving them complimentary feedback on their performance on a word-association test administered immediately prior to the "psi" word-association test. For ethical reasons the authors chose not to induce a negative self-concept in the remaining subjects but rather gave them no feedback on the first word-association test. The manipulation was found to have no significant effect on the psi scores.

In summary, six hypotheses based on propositions from the PMIR model and involving relationships to independent or predictor variables were tested. Three of the six were significantly confirmed, and in all six cases results were in the predicted direction.

Criticisms

Neither the PMIR model nor the research program surrounding it has been the object of critical review either inside or outside of parapsychology. Ironically, the one serious criticism directed to the model exclusively has been by Stanford himself. Stanford (1978) came to question the assumption, which the PMIR model shares with all traditional conceptualizations of psi, that "ESP at the most fundamental level is a form of communication" of information across a channel (p. 198). With respect to PK, he specifically questioned the assumption that PK is guided cybernetically by unconscious ESP (e.g., ESP must be used to monitor the ongoing status of the tumbling die so that PK can ultimately guide it to come to rest with the target face uppermost). Stanford labels these assumptions collectively as the "psychobiological model of psi...function" (p. 198).

Stanford based his questioning of the psychobiological model on research evidence suggesting that psi scores do not seem to be related to the complexity of the information source in ESP or the complexity of the target system in PK. For example, ESP performance does not seem to deteriorate if information from several sources must be integrated to make a response, and PK performance does not seem related to the complexity of an REG. In other words, the psychobiological model implies that as the requirements for cognitive processing capacity increase, psi performance should deteriorate. That does not seem to be the case.

Stanford thus chose to substitute the term conformance behavior for the term psi-mediated instrumental response. The new term became the foundation for what Stanford called his conformance model of psi. The new model

retains the dispositional assumptions of the PMIR model but eliminates the objectionable "psychobiological" assumptions. A novel feature of the model is the use of the REG as a general metaphor for the object of psi influence. Most notably, in the case of ESP the brain is conceptualized as an REG. The source of psi is a "disposed system" that influences the "REG" in such a way that an outcome is produced that is serving the needs of the disposed system. Thus, in ESP the brain is biased much like an REG to produce a cognition or behavior that serves the organism's needs. A particularly important corollary of the conformance model is that conformance is facilitated to the extent that the object of psi influence is labile; that is, that it exhibits properties of randomness or "free variability."

The relationship between the conformance model and the earlier PMIR model is not clearly stated but it would be reasonable for a reader to conclude that the conformance model is intended to replace the PMIR model. In any event, the PMIR research program was abandoned and Stanford no longer incorporates the PMIR model into his writings in a substantive manner. Although the conformance model has inspired some research both by Stanford and others (e.g., Braud, 1980), it has failed to generate the kind of systematic research program produced by the PMIR model.

Evaluation

Insofar as one is willing to allow paranormal constructs into scientific theorizing, I find little to criticize in the PMIR model per se. Its validity rests of course on its empirical track record, but the model appears to be internally consistent, its terms clearly defined and operationally definable. Its propositions are not expressed quantitatively, but this is true of most theorizing in psychology. In fact, given the poor reliability of psi measures it could be argued persuasively that any attempt at a quantitative theory or model at this stage of the field's development would be premature. Therefore, I will devote my evaluation primarily to the

research.

The word-association ESP test strikes me as basically sound, again assuming reasonable competence in its execution. The fact that subjects were not told they were taking an ESP test reduces considerably the possibility of subject fraud, at least insofar as members of the subject pool were not tipped off as to the true nature of the study by previous subjects. It appears that some debriefing sometimes took place at the end of the sessions, so the possibility of "leaks" cannot be entirely ruled out. Even so, the possibility of sensory cues was apparently eliminated by keeping the person administering the word-association test blind to the key stimulus word and the response contingency; i.e., it is unlikely that subjects could have cheated even if they had known the experimental hypothesis and were motivated to cheat.

It is not clear to what degree possible errors in recording the response latencies were precluded. In the Stanford and Thompson (1974) experiment the method is not described at all. In the subsequent experiments it is indicated that an electric timer was used, but it is not clear if the device automatically recorded the latency or whether this was done by hand. More importantly, it is not clear what steps were taken to assure uniformity across trials in the starting and stopping of the timer with respect to the subject's utterances. In at least one experiment (Stanford & Thompson, 1974), the recording was performed by the subjects themselves. However, even if the recording of latencies was not error-free, the fact that the tester was blind to the key stimulus word assured that there was no systematic bias in the recording; at worst, error variance was introduced.

The key stimulus word was chosen separately for each subject by means of a random number table. Although the exact method of target selection was not given, the fact that the number of potential key words was identical to the number of digits in the table minimizes the possibility of the kinds of

abuse uncovered in the Maimonides dream studies (see Chapter 2). Moreover, Stanford is one of the more careful psi researchers when it comes to these kinds of subtleties.

Finally, the methods of statistical analysis were appropriate and straightforward. Separate scores were computed for each subject and efforts were made to assure that the distributional assumptions of the parametric tests were met. Only in the case of the Stanford and Rust (1977) experiment did a problem arise as to which of two analyses of the same hypothesis was considered the primary one, and this does not affect the overall evaluation of the success of the research program one way or the other.

Most of my criticisms concern the procedures used to test the various hypotheses about psi. My major criticism in this connection is the lack of any checks to determine if the experimental manipulations had the desired effect. How do we know that the picture rating was pleasant and the pursuit-rotor task unpleasant? At a minimum, there likely were individual differences in subjects' responses to these tasks, especially to the picture-rating task, that could have been partialed out of the results. Even more problematic were the manipulations of need strength and self-concept. In fact, Stanford conceded retrospectively that one of his need-strength manipulations was unsatisfactory, based on informal comments by the subjects. Finally, it would have been relatively simple to check if short latencies to the stimulus words indeed were positively related to the choice of primary responses, as demanded by the "ready-responses" hypothesis.

Regarding the latter hypothesis, it will be recalled that it was tested in the Stanford and Stio (1976) experiment by manipulating whether the fastest or slowest response to the key word caused the subject to enter the "pleasant" condition. In this case the hypothesis was confirmed. However, this same manipulation was introduced in two other experiments but the results were never reported (Stanford & Thompson, 1974; Stanford &

Associates, 1976). Even though this manipulation was not designated in these studies as tests of formal hypotheses, they nonetheless bear upon the robustness of the finding by Stanford and Stio and should have been reported.

Although adequate randomization procedures were always used in choosing the key stimulus word, the same cannot always be said unequivocally for the assignment of subjects to experimental conditions. In the Stanford and Associates (1976) experiment, it is clearly stated that a random number table was used to assign the response contingency and an appropriate alternation method was used in the assignment of subjects to the self-concept conditions. A similar alternation method was also used in assigning subjects to experimenters in the PK experiment (Stanford et al., 1975). However, in the other cases the method of assigning subjects to conditions was not clearly specified, although it always seemed to involve some kind of randomization.

My final criticism concerns what I consider to be the premature abandonment of the PMIR research program. Although there is some merit to the argument that psi does not operate entirely in line with what would be expected by the information-processing assumptions of the psychobiological model, this is only one element of the PMIR model and is not necessarily the most important one. The assumptions about the unconscious and nonintentional nature of much psi functioning, as well as the psychodynamic assumptions, have not been challenged. Although the cybernetic guidance of PK indeed appears absurd after Stanford's analysis, it is less clear why the available evidence suggests so sweeping an abandonment of information-processing concepts as the conformance model seems to imply. ESP, at least, must at some stage interact with the cognitive processes of the brain if a meaningful response is to be elicited. The various cognitive mechanisms postulated in the PMIR model, such as the unconscious timing mechanism, need not be abandoned. They are not in fact inconsistent with the

conformance model, since any response, even conformance behavior, requires some kind of cognitive mediation at some stage. Stanford (1982, p. 19) acknowledges all this, but his use of the term "psychobiological" to label the model he proposes to replace implies a more radical revision than either logic or the data justify, or than he really intends.

It seems to me that a much better strategy would have been to modify the PMIR model rather than to abandon it in favor of a whole new model. The conformance model lacks all the conceptual elegance of its predessessor. The vagueness of its basic premise has led to much confusion and has triggered heated and unenlightening controversy about such things as whether the model is causal or whether it predicts that nonliving matter has psi (e.g., Beloff, 1979). The research it has inspired has been related almost exclusively to the corollary proposition of lability, which could have been attached to the PMIR model just as easily as to the conformance model. Stanford (1967) himself had introduced a very similar notion in the 1960s, which even antedated the PMIR model.

The PMIR model has been one of the most promising developments in parapsychology in the past 20 years. We can only hope that some day it will be resurrected, even if it must wear a slightly different wardrobe.

Chapter 9

METAL BENDING

Most of the PK research taken seriously by the parapsychological community is of the type exemplified by the REG experiments discussed in Chapter 5. This is often called micro-PK because the effects are of slight magnitude and require for their detection the application of statistical tests over a series of trials. In contrast, macro-PK refers to larger scale effects each of which is detectable by the naked eye. Many effects, i.e., single-trial effects detectable only by electronic amplification, fall in between these two extremes but are generally included under the heading of macro-PK.

Because of the rampant fraud associated with Spiritualist "physical mediums" of the 19th and early 20th centuries, macro-PK has been a taboo subject in parapsychology, especially in Britain and the United States. The recent revival of interest in macro-PK in general, and metal bending in particular, can be attributed to publicity surrounding the controversial Israeli psychic, Uri Geller. Perhaps the most important consequence of the Geller craze from the researcher's standpoint was that a number of less celebrated individuals, particularly children and teenagers, reportedly were able to bend metal after watching Geller do it. These "mini-Gellers" seemed to be a more promising research population than Geller himself, particularly since some of them appeared to be able to produce effects without touching the specimen.

The most extensive metal-bending research has been conducted by Dr. John Hasted, Professor of Experimental Physics at Birkbeck College, University of London. His most substantive work, which will be the focus of this review, has been published in five experimental reports in the Journal of the Society for Psychical Research (Hasted, 1976, 1977, 1978; Hasted & Robertson, 1979, 1980, 1981). This research, along with other generally

less formal work which includes such exotic fare as teleportation and levitation, is summarized in a book entitled The Metal-Benders (Hasted, 1981).

The research to be reviewed here involves protocols in which the subject was not allowed to touch the specimen. Exact procedures varied somewhat from session to session and procedures were almost never reported in precise detail. Nevertheless, certain general features can be described.

In the beginning of the research the specimens were latch keys, but in later research these were replaced by metal strips or bars, usually of aluminum or an aluminum alloy. The measuring instruments of primary interest were resistive strain gauges mounted either on the surface of the specimen or between layers of the metal sealed by epoxy resins. The strain gauges were connected by wires to a polygraph for amplication and recording of the signals. The wires were also used to mount the specimens; i.e., the specimens hung from the wires. The subject was seated in front of the specimen and generally allowed to point at it as long as the finger remained at least several inches away.

The primary control against the touching of the specimen was visual observation of the subject. However, since sessions often lasted up to two hours, Hasted conceded that full attention by the observer(s) throughout the period could not be maintained. Supplementary controls both against external physical fo ce and electrostatic or electromagnetic artifacts included: (a) electrode sensors designed to register touching of the metal, (b) electrical shielding of the strain gauges, (c) dummy loads, and (d) video recording of target strain gauges. None of these controls, except possibly the first, were utilized in all sessions, although anomalous phenomena were recorded in the presence of each. However, details about the implementation of the controls, e.g., the precise locations of the dummy loads, were rarely reported.

The subject population consisted primarily of middle-class British teenagers, frequently coming from families with academic backgrounds. According to The Metal Benders (p. 30), Hasted has achieved positive results with 20 subjects, although he claims to have worked extensively with only six of these. Two subjects, Nicholas Williams and Stephen North, both adolescents, contributed the vast majority of the data to be covered in this review.

Several generalities about the anomalous signals have been reported. The signals vary in strength from a few millivolts to a few volts and are generally two to three times the background noise. Compared to signals produced by physically touching the specimen, they have sharp peaks and short rise times (approx. 200 ms). It is not clear whether Hasted is claiming that signals with these characteristics cannot be reproduced by touching or whether typical touches do not have these characteristics. Permanent bends, which are reflected in baseline changes of the chart records of nearby strain gauges, sometimes are observed and sometimes are not.

Several specific experiments or, more precisely, groups of sessions using the same basic protocol, will now be summarized.

Basic Effects

Numerous signals were recorded from a strain gauge mounted on a latch key in a two-hour session with Nicholas Williams as subject (Hasted, 1976). A complete record of the chart tracing was published. Three successive, permanent bends of approximately 10, 50, and 12 degrees were determined by tracing onto paper. The last two of these apparently occurred after cessation of effort, but the report is not entirely clear on this point. It also would appear that the final "bend" was actually a restraightening.

"Synchronicity"

Sometimes, multiple specimens were employed at a single session, ranging in number from two to six. The configuration of the specimens varied from session to session. On the horizontal plane, they were either arranged "radially" from the subject (on a straight line outward from the subject), "equidistant" (subtending an angle of about 30 degrees as if on the rim of a circle with the subject at the center) or "opposite" (the subject is in between the specimens, one in front and one in back). In other sessions, one or more of the specimens in either the radial or equidistant horiztonal plane was displaced vertically with respect to the others.

Two major experiments with this procedure were reported. The first, with Nicholas Williams as subject, consisted of eight sessions and was limited to two or three specimens. The last session was videotaped (Hasted, 1977). It would appear from Hasted's diagrams that the sensors were at least one meter apart and the subject at least one meter from the closest sensor, except during the first session when he was somewhat free to move about.

It seems that a total of 54 signals appeared during the course of the experiment, of which 34 were designated as "synchronous" and 20 as "nonsynchronous." The classification was apparently made by visual inspection. Only in the equidistant, purely horizontal configuration did nonsynchronous signals seem to predominate. Although Hasted did not perform statistical tests, a chi-square test I performed comparing the proportion of synchronous signals in this configuration to the combined totals for the other configurations was significant. Permanent bends of the keys were detected in two of the sessions but the videotaped session was not among them.

The second experiment involved six sessions with Stephen North as subject, with the number of specimens increased to four, five, or six

(Hasted & Robertson, 1980). Adjacent specimens were apparently closer to each other than in the previous experiment. Apparently in contrast to the earlier sessions, dummy loads were uniformly applied as controls in these sessions. It would appear that a total of 66 signals were obtained. I could not determine precisely the proportion that were considered synchronous, but it seems to be comparable to that obtained in the previous experiment.

In both experiments, the proportion of synchronous signals appeared to be greatest with the radial-vertical configuration. This led Hasted to postulate a "surface of action" as a kind of vertically-oriented field extending outward from the subject.

Rotation

In a session with Nicholas Williams, Hasted (1977) took two strips of aluminum alloy, folded one around the other, and placed them on a table inside an empty room. Hasted and Williams waited outside. On this and subsequent occasions, one of the strips was later found to have been twisted around its vertical axis over part of its length. The effect only occurred when no one was watching. Hasted tentatively interpreted the effect as involving a rotation of the surface of action.

Extensions and Contractions

These studies were designed to determine whether the signals seemed to represent the kinds of forces necessary to produce bending. To detect this it was necessary to place sensors across the width of the specimen.

In a preliminary study of six sessions with three subjects, metal strips (mostly aluminum) were employed with sensors on the upper and lower surfaces (Hasted, 1981). The bars used were of different thicknesses, although thickness was not varied independently across subjects. A graph was printed indicating that the ratio of "bending" to "stretching" signals

decreased as the thickness of the specimen increased, but no supporting statistics were provided.

In a more elaborate study involving three sessions with Stephen North, six strain gauges were implanted across the width of an aluminum bar or in between strips of an eutetic alloy sealed together by epoxy resin (Hasted & Robertson, 1979). In both cases the four internal sensors were actually inside the specimen, not on the outer surface. To produce a bend, extension signals would need to be produced on one surface and contraction signals on the opposite surface, with the internal sensors expected to yield smaller signals of the same polarity as the external sensor closest to it.

However, the data revealed no such consistency. Of the 119 arrays recorded, only 17 (14%) corresponded to a simple bend or stretch pattern (no gradient changes). Most of the arrays had either one, two, or three gradient changes. In other words, the signals within the arrays seemed to distribute themselves randomly, as if they were independent of one another. Hasted labeled the effect "metal churning" in contrast to "metal bending." The effect seemed to imply that a visible bend can only occur on those relatively rare occasions when the forces across the width of the specimen happen to align themselves the "right" way. This of course is consistent with the observation that the number of signals detected on the chart recorder was much greater than the number of bends detected in Hasted's experiments.

Direction

In order to assess the direction of the forces operating on the surface of the metal, five sessions, four with North and one with another subject, were conducted using metal squares or discs instead of bars (Hasted & Robertson, 1979). A configuration of three strain gauges was set up on the surface of each specimen, two pointing orthogonally to each other and a third bisecting the angle between the other two. From these, separate

extension and contraction vectors could be calculated for each signal.

According to Hasted, the application of positive stress should produce extension along one diameter and an equal contraction along the opposite diameter. Again, however, the results were not as orderly as this hypothesis would lead one to expect. First of all, no preferred directions of strain could be detected. Moreover, there were no consistent ratios between the magnitudes of the corresponding extension and contraction signals. In fact, in about 25% of the cases extensions were accompanied by extensions or contractions by contractions: "metal churning" again.

Location

Two experiments were conducted to determine the localization of the ostensible strain along the length of a metal strip. In the first experiment, consisting of three sessions, three strain gauges were aligned along the surface of the strip (Hasted, 1978). In a later experiment, which involved five sessions with North as the subject, the number of sensors was increased to five (Hasted & Robertson, 1980). Dummy loads were also utilized in this latter experiment.

In both experiments the output tended to be greatest from the middle sensor. Hasted equated the distribution of signal strengths along the lengths of the specimen to a Gaussian distribution. Although it strikes me as problematic to define a curve by five and especially by three data points, it seems fair to say that the strength of the signals tended to fall off monotonically and symmetrically from the center.

Electrical Effects

Occasionally during the course of the research the electrodes which had been mounted for the purpose of detecting touches by the subject responded in the apparent absence of touch. It was subsequently found that North continued to be able to produce such artifacts without touch when a low

impedance operational amplifier supposedly immune to such effects was attached to the electrode. Only North seemed able to affect the apparatus in this way (Hasted & Robertson, 1981).

To further test for electrical effects, North participated in a series of seven sessions using two metal bars in a radial configuration. Both the electrodes and strain gauges were utilized as sensors. Almost half the signals (44%) registered exclusively on the electrodes, with 24% exclusively on the strain gauges and 32% on both. The proportion of electrode activations increased over sessions. Hasted speculated that North's awareness of the increasing interest in the electrode effects contributed to their increased prevalence.

In a subsequent series of ten sessions with North, an attempt was made to determine whether the effect was on the electrodes themselves or on the surrounding atmosphere. Two electrodes separated by distances ranging from 0.4 to 6.2 cm were given charges of +9V and -9V, respectively, the potentials being reversed every 11 seconds. Their hypothesis predicted that under the conditions of their experiment, if the signals were associated with atmospheric ionization charge bursts would appear uniformly at the oppositely charged electrode, whereas no such correlation would be found if the signals originated from the electrodes directly. It was found that 95.1% of the 1123 recorded signals behaved in accordance with their atmosphere-ionization hypothesis.

However, this conclusion was contradicted in yet another experiment (Hasted & Robertson, 1981). Hasted came to realize that previous results could be accounted for by assuming the origin of the charge to be on the subject's body and that it travelled through the atmosphere to the target along what he called a "temporary 'pranormal conduction' path" (p. 181). He reasoned that the atmosphere-ionization hypothesis would be refuted if it could be shown that a high-frequency signal could be transferred from a subject's body to the target. Such a signal could not be transmitted by

drift or diffusion, the base for the atmospheric-ionization hypothesis. Thus, a 10 kHz potential was transferred to Stephen North's body by placing close to him a 10 kHz oscillator connected to a metal plate or "antenna." As predicted by the "conduction path" hypothesis, the 10 kHz signal was also momentarily transferred to or induced on a partially screened electrode in the vicinity of North. This effect was not obtained with control subjects.

Piezoelectric Sensors

In the most recent phase of his research, Hasted has shifted from strain gauges to piezoelectric sensors (Hasted, Robertson, & Arathoon, 1983). As used by Hasted, piezoelectric sensors measure the rate of change of stress rather than the level of stress per se. This makes them more sensitive than the strain gauges to the rapidly varying pulses that seem to characterize the ostensible PK effects. However, in order to minimize electrostatic artifact, Hasted had to eliminate much of this added sensitivity by connecting the high resistance piezoelectric transducer across a relatively low resistance (3.5 K ohms). Nonetheless, the overall piezoelectric system was still more sensitive than the strain gauges to the signals of interest.

Hasted briefly reported results from eight sessions with Stephen North and two other subjects. The sessions with North and one of the other subjects were held in an electrically shielded room. Control against touching or fraud continued to be through observation by the experimenters and (in some cases) observers. A dummy channel (unscreened input resistance) was situated somewhere inside the screened room and connected to a separate amplifier and recorder as a check for electrical artifacts originating inside the room. None were found.

In most of the sessions, at least, a strain gauge and piezoelectric sensor were mounted back to back on a thin metal specimen. In the first session (with North as subject), 15 signals appeared on the piezo channel

and only nine on the strain-gauge channel. Only four of these signals were synchronous. In subsequent sessions, both with North and the other subjects, there were virtually no signals on the strain-gauge channel, whereas the density of signals on the piezo channel remained about the same on the average.

Although Hasted speculated that this change may have had something to do with an increase in sensitivity of the piezo channel following the first session due to improvements in the electronics, it is difficult to see how this could account for the lack of strain-gauge signals.

Hasted also reported the ability of Stephen North to exert some control over the timing of the signals in this phase of the research, but the details were sketchy.

Criticisms

To this reviewer's knowledge, no comprehensive critiques of Hasted's research have yet been published. Perhaps the closest approximation is a review of The Metal Benders by Stokes (1982). Wood (1982) raised technical objections to the interpretations Hasted placed upon his "strain" signals, expressing particular concern about their small magnitude. He also questioned Hasted's assumption that the extension and contraction vectors should be equal for the metal disc experiment (p. 184), and he noted that the rotation effect (p. 182) could be produced normally because such twists are caused by shear rather than by extension forces. Hasted had assumed the latter in arguing for the effect being paranormal. Hasted replied to Wood's criticisms in the same article.

Also worthy of mention at this point are brief comments by an electronics expert named Horowitz (cited by Randi, 1982), who maintained, apparently rather indignantly, that the signals which appeared on the chart recorder in Hasted's experiment are readily explicable as electrical transients picked up by the amplifiers. Although this criticism is clearly

applicable to Hasted's earlier work (which may have been all Horowitz had access to), its applicability to the later work in which a dummy load was utilized is less certain.

A thorough, albeit sympathetic, critique of Hasted's experiments has appeared in an unpublished doctoral dissertation by Isaacs (1984). A particularly valuable aspect of this review is that Isaacs obtained information directly from Hasted about certain procedural details which did not appear in the latter's published reports.

These critics' points are generally included among those arrived at independently by myself and my consultant. Therefore, I will leave further discussion of them to the following evaluation section.

Evaluation

The first general question to be addressed about Hasted's work is whether the effects he has reported can be attributed to normal, i.e., artifactual processes. This question must be addressed separately for the gross metal-bending (deformation) effects and the more subtle effects detected on the chart recorders.

Deformations

Because of the physical setup, it is hard to imagine how the subjects could have physically bent the specimens while they were attached to the recording devices without detection by an experimenter (or, the video recording, when used), or without leaving an obvious tell-tale trace on the chart record. This comment does not apply to the twisted metal strips, however, which were left unobserved in a room. In this case, documentation is insufficient to rule out someone entering the room undetected and manipulating the specimen. Although twists as tight as those observed seem difficult to produce, even granting that shear forces are involved, the difficulty or possibility of mechanically producing such deformations cannot

be assessed without extensive control tests.

In none of the cases is information given to reassure the reader that either physical deformation of the specimens or substitution of an already deformed specimen was precluded as a possibility at some point during the session (e.g., before the specimen was mounted). In particular, I could find no mention of specimens having been marked. Although no positive evidence of such manipulations exists, Hasted's lack of sensitivity to this issue in his reports reduces the confidence one can place in the observed deformations being truly anomalous. The fact that his subjects were teenagers is not an argument against trickery being employed, although Hasted sometimes implies that it is.

Chart-Record Signals

The signals on the chart records could in principle be produced artifactually either by direct interaction with the specimen (or the sensor(s) attached to it) or interaction with the peripheral devices (i.e., amplifiers, chart recorder, etc.). Possibilities for direct interaction with specimen and sensor include touch, air currents (e.g., blowing on the specimen), auditory stimuli (e.g., ultrasonic sounds), thermal stimuli, and localized electrical signals.

Even granted the unreliability of long periods of human observation, it seems unlikely that a subject could consistently get away with touching a specimen without being detected. This is especially true in the case of Nicholas Williams, who customarily stationed himself several feet from the specimens. Also, it again should be noted that some sessions were videotaped, and touch detectors were sometimes employed. Blowing on the specimens would be more difficult to detect, however. According to Isaacs (1984), only air currents powerful enough to cause the rigidly mounted specimens to swing would be powerful enough to be detected by the amplication and recording system. However, Isaacs stopped short of saying

that such effects are impossible, and no control trials have been reported to assess this potential artifact systematically. Issacs maintained that Hasted's amplification and filtering system would have precluded auditory effects from being recorded. In some sessions, Hasted controlled against thermal effects, which could also produce air currents, by employing thermal sensors.

Although one can imagine many sources of gross electrostatic or electromagnetic artifacts, localizing them to a particular specimen is a different matter. However, as Hasted recognized, it is possible that a strain gauge could be triggered either by the subject building up an electrostatic charge in his body and moving a finger, say, close to the specimen, or by creating dynamic electrostatic induction through gross body movements. On the other hand, such potential effects, even if the requisite movements had escaped visual detection by the experimenter(s), would have needed to overcome the electrical shielding of specimens routinely applied in Hasted's later work. Electrostatic effects should also have been picked up by the touch detectors; the problem here, of course, is that on some occasions these detectors were triggered, and some of the anomalous chart recordings are now conceded to have been electrical in origin. For the reasons cited above, it is unlikely that all these triggerings of the touch detectors can be attributed to undetected touch, but what they are attributable to remains uncertain.

Another argument against hypotheses based upon localized artifacts is the frequent occurrence of "synchronous" signals associated with sensors located up to several feet apart. The problem is that the signals could conceivably radiate out from the vicinity of one sensor to another, even over the distances of separation utilized. If the signals were truly synchronous, this hypothesis might be precluded. However, Hasted's recording mechanism was not adequate to define synchronicity with the necessary precision; i.e., the diagnostic equipment was too slow to measure

the time it would take for the radiation to propagate. Hasted's "synchronous" signals can only be considered synchronous in a loose sense of the term.

The main control against global artifacts (and indeed the most important control in all of Hasted's work) was the use of a dummy load with its own amplifier. The dummy load would be expected to pick up gross electrical artifacts due to the switching on and off of appliances, etc., as well as most signals from simple devices that might be smuggled into the laboratory by a subject with intentions of fraud. However, no data were reported comparing empirically the effects of various signals on the dummy loads and the strain gauges. Such data would have made this control more reassuring.

The remaining potential source of artifact in this category is direct interaction with the chart recorder or the chart recorder pens (Randi, 1975). Hasted (1981) claims, however, that the equipment was always kept well out of the subject's reach.

In conclusion, assuming normal experimental competence and honesty, it appears unlikely but possible that the effects on Hasted's chart records can be explained away as mundane artifacts.

Process-Oriented Data

The second general question to be addressed in evaluating Hasted's research is of a more process-oriented character; namely, what can be said about the mechanism underlying the effects, assuming they are not mere artifacts as discussed above?

Unfortunately, as also was the case in the REG research (Chapter 5), Hasted's research methods do not lend themselves well to drawing conclusions of a process-oriented nature. Three distinct classes of deficiencies can be cited:

(1) Methods of recording the anomalous signals were not well suited to providing precise characterizations of them. Most importantly, the chart recorders Hasted used were too slow (approximately .1 sec) to record reliably the rapidly rising signals of primary interest, which could have resulted in the loss of data. In general, signals were not recorded or reported in such a way as to allow confident determination of their nature. The problem is not so much the strength of the signals, as suggested by Wood (1982), but rather their qualitative characterization.

(2) Principles of good experimental design were largely ignored. One never finds systematic comparisons of experimental and control conditions in Hasted's work. Successive tasks were not counterbalanced to eliminate possible order effects. Most importantly, potential psychological and physical effects were continually confounded. No efforts were made to keep subjects blind to experimental manipulations and hypotheses, thus making it impossible to distinguish basic physical characteristics of the phenomena from characteristics associated with, and thus constrained to, the psychological needs, attitudes, and intentions of the subjects. As different tasks were given to different subjects, it is difficult to assess the generality of the process-oriented data obtained or to make proper between-subject comparisons.

(3) Although many of the effects upon which conclusions were based did not occur consistently, the conclusions were not backed up by the requisite statistical analyses, a point also stressed by Stokes (1982).

Examples of how these deficiencies contribute to ambiguity in the interpretation of Hasted's results will now be given. Perhaps the most important of these examples concerns the basic nature of the recorded signals. Although the signals are often referred to in passing as reflecting strain (i.e., extension, contraction, or bending of the metal specimen), only rarely is such a conclusion justified. This is true even if we agree that the signals are not artifactual in origin. Hasted's more

recent work has indeed illustrated that many of the signals in that work seem to be electrical in nature, which raises the possibility that some of the effects in the earlier work may also have been electrical. Only in those sessions where the signals were shown to conform to permanent deformations of the specimen does the case for their representing actual strain effects appear to be strong (Hasted, 1977). The incapacity to characterize the remaining signals is attributable in part to the suboptimal recording techniques mentioned above.

Hasted's conclusions regarding the "surface of action" are also problematic, even if one accepts the loose definition of "synchronous." This concept is based on observations of data from North and Williams that synchronous signals are more prevalent when the specimens are in a radial-vertical configuration with respect to the subject than in some other configuration. However, no statistical analyses were offered to support the significance of this trend. In the case of Williams' data (Hasted, 1977), 12 of 15 (80%) signals in the vertical or radial-horizontal-vertical configurations were synchronous as compared to 22 of 39 (56%) with the other configurations. This difference is associated with a corrected chi-square value of 1.67, which with one degree of freedom is clearly nonsignificant. The trend in North's data seems somewhat stronger (Hasted & Robertson, 1980), but I was unable to perform a statistical analysis from the available data. Since the order of presentation of tasks to the subjects was not counterbalanced, the trends that were uncovered might be due to order effects (e.g., a subject might do relatively well with a particular configuration simply because it was presented in an early or late session). Although the surface of action is presented as a basic physical characteristic of the phenomenon, it could just as easily be a reflection of a possible psychological preference of North and Williams; there is certainly no basis for drawing conclusions about the generality of the surface of action.

Finally, there are problems with the concept itself. Since it is little more than a metaphor for certain empirical observations, its theoretical value is limited; it does not seem to be, or at least has not been shown to be, a source of hypotheses or predictions that might increase our insight into the mechanisms involved. Also, as noted by Stokes (1982), subsequent assumptions about the surface of action moving outward from the subject (Hasted & Robertson, 1981) seem to contradict the data that signals appear synchronously on sensors at different radial distances form the subject. The willingness to add on assumptions about the movement of the surface of action threatens to render the concept unfalsifiable.

The data concerning the localization and direction of the ostensible psychokinetic forces and whether they are extensions or contractions suffer from comparable ambiguities to those addressed in the preceding paragraphs. As noted by Wood (1982) regarding the experiments with the metal discs, it is not always clear what a strain hypothesis would predict. For instance, one would not expect symmetries of the type postulated by Hasted to be found if the strain were localized on particular sensors. However, even if one were justified in adopting, say, a simple bending hypothesis, failure to confirm it could not be interpeted. As noted previously, the signals which could confirm such a hypothesis might never have been registered due to deficiencies in the recording procedure. This recording problem could have added noise to virtually all of Hasted's process-oriented data. On the other hand, the failure to detect regularities in support of a strain hypothesis, for whatever reason, reinforces concern that the signals may not be strain signals at all.

Finally, the nature of the "electrical effects" reported by Hasted is still not clearly resolved. The experiment supporting the "paranormal conduction path" is not sufficient by itself to settle the matter. In particular, it is unclear to what extent the results of this experiment can be generalized to the test situations in Hasted's other experiments.

In conclusion, Hasted has presented us with a set of intriguing anomalies about which we can say little, despite his generally process-oriented approach. Much of the problem may be due to the quality of his research reports, which I consider the most deficient of any considered in this review.

As noted at the beginning of this chapter, this review has been limited to the better controlled of the effects reported by Hasted. Particularly in his book, Hasted (1981) intersperses accounts of these experiments with more informal observations. These commentaries sometimes project an aura of credulity which has been alluded to by sympathetic (Collins & Pinch, 1982), unsympathetic (Randi, 1982), and neutral (Stokes, 1982) reviewers of his work. This is particularly true of the chapter in his book devoted to informal (and often poor) observations or inferences of ostensible teleportation, often in the vicinity of Uri Geller. When a scientist writing a scientific book starts describing the teleportation of a liver from his Christmas turkey, even his more controlled observations are likely to lose credibility in the eyes of many scientists. Hasted (1976) has defended the presentation of such material and even maintains that "one's own credibility is relatively unimportant" (p. 382). However, the problem as I see it is not the reporting of anomalous phenomena per se (even very anomalous phenomena), but rather the according of evidential weight (by implication, at least) to poorly controlled observations. At a minimum, I think Hasted has used poor judgment in his presentation of this material. Although I personally am less willing than many scientists to draw inferences about the reliability of experimental reports from such ad hominem considerations, I do feel they should be placed on the table for the benefit of readers who might think otherwise.

Structural Changes

As a general rule, metal-bending experiments are only considered valid if protocols are used which prevent the subject from touching the specimen. However, experiments which allow touch are sometimes considered worthy of serious attention if it can be shown that structural changes in the metal were produced that are inconsistent with a physical-bending hypothesis.

The most impressive results of this type have been provided by experiments conducted by the French metallurgists Charles Crussard and J. Bouvaist (1978) with a subject named Jean-Paul Girard. During the course of the research Girard deformed or transformed 150 specimens, but the authors considered only 20 of these episodes to be evidential. Eight of the 20 were selected for detailed review in their report.

The exact procedures varied from trial to trial, but in most cases it involved Girard holding a metal bar first outside and then inside a stopped glass tube. Observations to determine whether the specimen had been bent were made before the trial, before the specimen was placed in the tube, and after the trial. Structural analyses were generally performed before and after the trial. Finally, simulations sometimes were performed on control specimens to assess what structural changes occur when deformations are produced by normal physical means.

The first two specimens were bars of aluminum alloys. In each case, bends of the bar were observed. The second bar was submitted to laboratory tests which confirmed that the force needed to bend it physically was twice that exerted by the strongest man who had previously attempted to bend the bar with his hands.

The second two specimens were stainless steel cylinders 7 mm in diameter and 85 mm long. Visible bends were detected only in the first specimen. However, both bars exhibited local magnetism that was not exhibited prior to the trial. Various other tests revealed the presence of martensite in the magnetic zones. The conversion of austenite to

martensite, which accounted for the magnetism, was not of a type that results from heating or cooling, but rather from deformation. However, the quantity of martensite in both cases was considered to be far greater than expected, given the degree of bending observed. The localization of the martensite was also considered surprising. Physical bending (30 degrees back and forth) of a control specimen from the same batch as the test samples caused the specimen to assume an S-shape which had not been observed with either of the test specimens. Also, the distribution of the magnetism in the test bar differed from that expected and observed in the control.

The final four specimens were Duralumin plates which Girard was asked to "compact," i.e., to harden. Although bending was observed in only one of the four specimens, measures with a Vickers microdurometer revealed increased hardening in all four cases, varying from 6% to 12%. The result for the fourth specimen was independently confirmed in Hasted's laboratory. Further tests revealed that the changes in hardness were associated with a modification of the residual longitudinal stress in the hardened zones. Moreover, tests of the first two specimens revealed an anomalous microstructure in the hardened zones consisting of a high density of small (200 angstroms) dislocation loops. It is not clear whether these deformations were not found in the remaining two samples or not tested for.

Several simulation tests were also performed with control samples. Achieving the degree of hardening observed in the test specimens by bending back and forth was shown to require a permanent deformation greater than what was actually observed in any of the test specimens. Also, the physical bending did not produce the small dislocation loops. A local compression test duplicated the requisite hardness but again without producing the dislocation loops. Also, when the equivalent degree of hardness was physically generated, the thickness of the plate was reduced 13%, compared to a maximum of 2% in the plates handled by Girard. The most successful simulation was produced by shot-peening, which duplicated the essential

features of the test specimens except that it dulled the surface. However, the surface shine could be restored by polishing.

The authors exhibited a commendable degree of caution in interpreting the results of their experiments. Labeling the effects as "abnormal" rather than "paranormal," they asserted that while their data "...rule out for the moment any explanation by known physical mechanisms or by tricks," they acknowledged the possibility that "a more insightful investigator may conceive a mechanism of which we did not think" (p. 13). They concluded that to overcome their controls, Girard "would have to be not only an accomplished illusionist...but also a first-class metallurgist." (p. 13).

Evaluation

As frankly noted by the authors, the results with Girard must be interpreted in light of the fact that he possesses conjuring skills; it turns out that he was even enrolled in the "Magician's Register." It is to Girard's credit, however, that he apparently volunteered to the authors during the course of the investigation the fact that "he had practiced prestidigitation" (p. 14). Less reassuring is the assertion by one of several magicians consulted by the authors that he discovered "a sign of trickery in a film J. P. Girard had obtained for us without telling us it was faked" (p. 14).

If trickery was used to obtain the effects, it seems unlikely that it involved physical bending of the specimen by Girard during the trial itself. All the trials were conducted in the presence of multiple witnesses and at least two of the trials were filmed. The deformations that were observed would require more force than that producible by a normal human. Finally, and most importantly, the structural changes found in most of the specimens did not conform to what would be expected by simple bending.

If the results are fraudulent, a more likely possibility in my view is that they were achieved by Girard substituting a previously doctored

specimen for the test specimen. The primary control the authors provided against this possibility was to mark the specimens. The methods used to mark the two aluminum bars was not specified. The stainless steel bars were marked with an electric pencil and the Duralumin plates with some kind of "iron".

To defeat these controls by substitution, Girard would have needed to take the following steps:

(1) Know in advance the type of specimen to be employed, obtain one or more examples of the specimen, and perform the necessary transformations.

(2) Duplicate on this specimen or specimens the markings made on the test specimen (after knowing the latter).

(3) Substitute his own specimen(s) for the test specimen at some point(s) during the trial.

Too little information is given in the report to allow the first step to be precluded. In particular, we lack information as to how much access Girard or an accomplice may have had to the labs or lab facilities at times other than the experimental sessions. We also lack sufficient information concerning how far in advance Girard knew the types of specimens that were to be employed at a given session. We do know in the case of the stainless steel cylinders, however, that Girard had the unmarked test specimens in his possession for several days prior to the trials.

Regarding the second step, the opportunity for Girard to mark the duplicate specimen is greater the longer the time interval between the marking of the specimen, or (more importantly) Girard's knowledge of the marking, and the trial. In the case of the second aluminum bar and the four Duralumin plates, we are not told when the markings were made. We are told that the stainless steel cylinders were marked at the beginning of the session, which certainly would restrict Girard's opportunities to mark duplicate specimens, but not necessarily preclude them. Such opportunities would appear to have been precluded unequivocally in the case of the first

aluminum bar, however, since it was marked immediately before the trial. This also was the one specimen that Girard was not allowed to touch at any time during the trial prior to verification of the deformation.

In the case of the second aluminum bar, distinctive flaws in the structure of the bar were cited as serving the same function as the markings. However, it is not clear how distinctive these flaws really were and whether they could also have appeared in other bars from the same batch.

The control against the third step (i.e., actual substitution) was the filming referred to earlier. However, I would not want to conclude unequivocally that a skilled illusionist could not overcome this control, especially in the absence of detailed information about how the filming was done.

Yet another possible mechanism for trickery is for Girard to have performed transformations on the test specimen itself prior to the session. This hypothesis seems precluded in the cases of the first aluminum bar and the steel cylinders, since tests for deformation and (in the latter case) magnetism were performed at the beginning of the session. The hypothesis also seems precluded for the second aluminum bar and first Duralumin plate, since successive deformations were observed during the course of the trial. No information is provided in the report that would preclude this possibility for the other Duralumin plates.

The evidential weight of the findings that emerged from the structural analyses per se is less than might have been hoped for. The effects on the second aluminum bar (the first was not evaluated this way) seem to conform to what would be expected by the application of ordinary physical force, even though the force is greater than that which could be generated by an unaided human. The authors concede that the effect on the Duralumin plates could have been duplicated by a combination of shot-peening and polishing. Only the structural changes in the steel cylinders seem truly anomalous, in that the amount and distribution of the martensite in the tests samples was

different than expected given the amount of permanent deformation observed. However, this conclusion rests heavily on the outcome of one control sample. Is it safe to generalize widely from this one outcome, or might a larger sample of such control specimens yield a distribution of outcomes which might include (albeit as a distinct minority) an outcome analogous to that found with the test specimens? The fact that the authors had to resort to a control sample at all suggests that the answer to this question may not be known.

Based on the information available in report, all the results except those with the first, and possibly the second, aluminum bars are potentially explainable by some form of substitution or pretransformation of specimens. We must also remember that the subject is known to have conjuring skills and might have used them at least once in the context of the research. On the other hand, several factors weigh in favor of the results being truly anomalous. As I argued in my discussion of the Delmore research (Chapter 6), the fact that a subject possesses conjuring skills is not by itself sufficient grounds for discounting evidence obtained with that subject under well-controlled conditions. The authors were aware of the possibility of fraud, consulted with magicians, and in fact took several precautions to preclude fraud. As a result, alternative hypotheses needed to account for the results by trickery seem unparsimonious. Finally, the straightforward results with the first aluminum bar seem especially resistant to normal explanation. For these reasons, it is my opinion that the modest conclusions reached by the authors are justified, and the "abnormal" results they have uncovered deserve to be taken seriously. Assuming that no more credible explanations of the Crussard results are forthcoming, further research with Girard is warranted (despite what independent evidence there may be of his use of conjuring) as well as a search for other subjects who might be able to produce comparable effects.

Chapter 10

SUMMARY AND CONCLUSIONS

This review began with the premise that no adequate critique of the experimental evidence for psi is possible without first addressing philosophical issues about how research questions in parapsychology have been formulated. In Chapter 1 it was maintained that the traditional demand of a "conclusive experiment" as a necessary condition to verify the existence of psi is inappropriate because it is inherently unfalsifiable. Replicability is also inadequate as a criterion, because the replicability of an effect says nothing about its cause.

The problem was pursued further by critiquing the formulation of parapsychology's fundamental question; i.e., "Does psi exist?" This question both reflects and reinforces the conflation of "psi" as subject matter and "psi" as explanatory principle. A more appropriate question is then proposed: "How can ostensible psychic events (OPEs) be best explained?"

This new question has several important implications. First, it implies that in order to demonstrate "psi" as a paranormal principle, researchers must empirically confirm a theory or model embodying such a principle, something parapsychologists themselves concede they have been unable to do. Therefore, the conclusion that parapsychologists have established "psi" can be rejected on logical grounds prior to evaluation of the data per se.

On the other hand, absence of an adequate paranormal explanation of OPEs does not imply the presence of an adequate conventional explanation. A second implication of the new question is that the burden of proof falls on <u>anyone</u> who claims to have explained OPEs either paranormally or conventionally. Thus, the important question to be addressed by examination of the empirical evidence is whether any conventional explanations of OPEs

can be considered scientifically adequate. The standards for evaluating conventional explanations of OPEs are the same empirical standards used in the rest of science. For present purposes, these include internal evidence for the hypothesis from within the experiment itself, empirical support for the hypothesis in related contexts, and the plausibility of the hypothesis.

The remaining chapters have critically reviewed ten major parapsychological research programs, eight from single laboratories and two from multiple laboratories, employing the perspective described above. Following is a brief summary of each of those reviews.

The Maimonides Dream Experiments

The first set of free-response ESP experiments to be touted as providing strong support for the existence of psi was a series of experiments on ESP in dreams supervised by Drs. Montague Ullman and Stanley Krippner at Maimonides Medical Center. Each night, a subject was awakened by the experimenter each time physiological monitors of his brain waves and eye movement activity suggested that he had been dreaming. The subject then was asked to give a dream report. Meanwhile, an agent located in another room periodically attempted to telepathically influence the subject's dream content by concentrating on a randomly selected art print. At the completion of each series, outside judges attempted to match up the dream reports (generally supplemented by the subject's morning-after associations to his taped dream reports plus a "guess for the night") to the targets on a blind basis. In most cases, the subjects also served as judges.

Eleven formal series were defined for the review, of which three involved one trial per subject and the rest, multiple trials per subjects. Two of the former group of studies were screening experiments used to select subjects for the more definitive latter group of studies. In addition, the results of several hundred pilot sessions were reported.

Using a crude method of analysis based on how frequently the judges ranked the target in the top half of the distribution, the combined results for the multiple-trial-per-subject series and the pooled results of the pilot sessions were both significantly positive, whereas the combined results for the single-trial-per-subject series were not.

Attempts to replicate two of the significant studies by noted dream researcher Dr. David Foulkes at the University of Wyoming, in consultation with the Maimonides team, both failed. Critic Hansel attributed the failure to tighter controls against fraud in the Wyoming experiments, whereas parapsychologist Van de Castle, one of the subjects in both the Maimonides and Wyoming experiments, stressed the debilitating effect of the skeptical attitude of the Wyoming team.

Other criticisms of the Maimonides experiments included the claim that the experimenter was not blind to the target, lack of a baseline or control condition, and possible lack of intrajudge independence of the ratings or rankings within a series.

The evaluation began with a meta-analysis of the eleven formal Maimonides experiments using, where necessary, a worst-case approximation of the variance to allow for the dependency problem. The suggestion that the experimenter was not blind to the targets was attributed to a misreading by Hansel of an ambiguous passage in one of the experimental reports. Control judgings were considered to be unnecessary because, for each trial, the other trials in the series served as an internal control. However, a previously undiscovered flaw in the way the targets were randomized in several of the significant Maimonides experiments and <u>possibly</u> in the Wyoming replications was revealed. However, for this flaw to have had practical consequences, certain other assumptions, the status of which is indeterminate, must be satisfied. They include such things as whether the judges knew the original order of the targets in the pool and whether this original order was determined randomly. Also, this criticism does not apply

to all the successful series. As an explanation for the overall results, this artifact is possible but highly improbable.

Remote Viewing

Several free-response ESP experiments using the remote viewing (RV) procedure have achieved considerable publicity. The procedure was originated by physicists Harold Puthoff and Russell Targ at SRI International. In the initial series, with geographical sites as targets, significant results were achieved by a group of nine subjects, with two subjects, Pat Price and Hella Hammid, making the most substantial contributions. Significant results were also obtained by a partly overlapping sample of five subjects in a subsequent series where the targets were pieces of office and lab equipment.

The main critics of the SRI experiments have been psychologists Richard Marks and David Kammann. Their primary criticism was that failure to edit the response transcripts combined with failure to randomize the materials given to the judge allowed the judge to infer which transcript went with which target site irrespective of the real accuracy of the subject's description. An exchange of correspondence by the protagonists, which came to include psychologists Charles Tart and Robert Morris as well, resolved that the criticism was applicable to Price's data but may not have been applicable to Hammid's data; however, other sources of biasing information may have been present in the latter case. Rejudging of the Price data with the biasing cues presumably removed from the transcripts confirmed the criticism when the rejudging was conducted under the auspices of Marks and Kammann, and refuted the criticism when it was conducted under the auspices of Tart, Puthoff, and Targ. Unfortunately, neither of the rejudgings was entirely adequate methodologically.

A somewhat related criticism was proposed by psychologist Ray Hyman. Given that the targets in a given series were sampled without replacement

and the subject received feedback after each trial, the subject could have gained unfair advantage by not including descriptors characteristic of targets in preceding trials in his responses on subsequent trials. The researchers countered that the criticism did not apply because redundancy of target characteristics was introduced into the target pool. This is an adequate rebuttal in theory, but more information about the targets would be needed to confidently assess its adequacy in practice.

A second major criticism by Marks and Kammann (part of which was later retracted) concerned circumstantial evidence that unsuccessful trials were either not reported by the SRI researchers or classified post hoc as informal. Analysis reveals that much of their case is attributable to self-serving and implausible interpretations of statements made by Puthoff and Targ, yet ambiguities still remain.

Major series of successful replications of the RV experiments have been reported by John Bisaha and Brenda Dunne and by Marilyn Schlitz. Although these experiments were methodologically superior to the earlier SRI studies, only the final Schlitz experiment seems to have fully addressed the sensory-cue criticisms of Marks, Kammann, and Hyman. This study achieved only modest significance. Unsuccessful replications have been reported by Marks and Kammann and by another researcher skeptical of psi, Edward Karnes. These unsuccessful studies themselves contained methodological flaws, although in the case of Marks and Kammann this could have been motivated by a desire to duplicate faithfully the SRI procedure.

Success at obtaining significant RV results thus seems highly correlated with the attitudes toward psi of the principal investigators. Differences in the motivation and enthusiasm of the judges selected by these investigators is suggested as one possible explanation of this finding. Finally, the reluctance of the SRI team to share their data with critical investigators is seen as damaging to their credibility, although it is perhaps understandable in light of the precedent set by some critics to use

such data as a basis for unsubstantiated insinuations of fraud.

The Ganzfeld Debate

Several major parapsychological laboratories have reported significant results in free-response ESP experiments in which subjects are exposed to a short-term perceptual deprivation procedure called the ganzfeld. A data base of 42 such experiments from ten principal investigators was the subject of an exchange between parapsychologist Charles Honorton and psychologist Ray Hyman.

The first issue addressed by the protagonists was the validity of the claimed 55% replication rate of the ganzfeld paradigm. Hyman made two kinds of arguments: first, the claimed success rate was too high because (1) experimental cells which should be treated as separate experiments were either pooled or arbitrarily excluded, and (2) unpublished failures, particularly if they involved small sample sizes, likely went unreported. Second, he claimed that the .05 level as a criterion for significance was too low because the published significance levels were not corrected for various types of multiple testing. Honorton's principal rebuttal consisted of an analysis of 28 of the 42 experiments for which a uniform test of significance could be applied and which yielded a highly significant outcome. He also noted that over 400 nonsignificant and unpublished studies would be necessary to reduce the overall data base to nonsignificance.

My evaluation supported the claim that the data base was of a non-chance character. Selection of the unit of analysis is somewhat arbitrary and choosing the cell as the unit would be expected to yield a lower success rate in the kind of analysis Hyman employed simply as a result of reduced power. Honorton's new analyses seem to adequately address the problem of multiple testing; moreover, the uncorrected p-values are the appropriate ones to use when the purpose is to assess experiments jointly as opposed to individually. Finally, Hyman failed to address the fact that the

great majority of the studies yielded results in the same (positive) direction.

The second major issue addressed by the protagonists was the methodological adequacy of the experiments in the data base. Hyman cited six categories of flaws each of which applied to 24%-74% of the studies. They included: failure to use duplicate target sets for judging, inadequate randomization of targets, inadequate randomization of judging materials, inadequate documentation, inadequate security against cheating by subjects, and improper statistics. He concluded that the pervasiveness of these flaws indicated general sloppiness in the conduct of the experiments. He undertook several multivariate analyses to show that these flaws could collectively account for the apparent significance of the results and could explain why some experimenters achieved more successful results in ganzfeld experiments than other experimenters. Honorton acknowledged the presence of the flaws but argued that Hyman exaggerated their pervasiveness by coding many studies as flawed that failed to meet his (Hyman's) stated flaw criteria. Honorton's statistical consultant, Dr. David Saunders, questioned the validity of Hyman's multivariate analyses, primarily on the basis of probability pyramiding, statistical dependencies among the variables, and insufficient sample size.

My statistical consultant agreed with Saunders' critique of Hyman's multivariate analyses. Moreover, the significant bivariate correlations between Hyman's flaw codings and study outcomes were attributable to coding errors on Hyman's part. The one exception was that studies which achieved target randomization by shuffling or similar informal methods did achieve significantly better outcomes than those using more reliable randomization procedures. In general, it was concluded that Hyman had not provided empirical evidence of a link between methodological flaws and study outcomes.

On the other hand, the fact remains that the flaws must still be considered in evaluating the data base. As Hyman points out, a flaw that fails to discriminate successful from unsuccessful studies might still exert a causal influence by interacting with other procedural factors. The various flaws were analyzed in terms of the plausibility of the scenarios they imply and found to be of questionable plausibility. In one case (failure to use duplicate target sets) empirical research of relevance to this issue exists. Given the objectives of most of these experiments, it is my opinion that generalized sloppiness cannot be inferred from the flaws that were uncovered, especially since most are reducible to incomplete documentation in the reports.

Random-Event-Generator Research

The pioneer of, and the major contributor to, systematic research with random event generators in parapsychology has been Helmut Schmidt of the Mind Science Foundation. His 15 years of research with REGs have yielded 14 experimental reports and can be divided into four phases. In the first phase, ESP methodology predominated, with the subject guessing which of four states the REG would select for each trial. In the second phase, subjects attempted to bias a rapidly generated sequence of events by PK. In the third phase, a version of the "observational theories" was tested by having subjects observe or listen to prerecorded sequences of targets and thereby attempt to bias them retroactively. In the fourth phase, the targets were pseudo-random sequences derived by applying an algorithm to a seed number generated by an REG.

Highly significant results in the direction of the subject's intent were manifested consistently in all phases of Schmidt's research program. The modest experimental manipulations Schmidt employed generally had little effect on the results. Type or rate of feedback, and whether the targets were prerecorded or contemporaneous, had no discernable effect at all.

There were some indications that slower target generation rates were associated with higher scoring and that actual effort to influence the REG, as opposed to just observing the feedback, facilitated scoring but was not necessary for PK to occur. However, since these variables were not manipulated in proper experimental designs, interpretations must be made cautiously.

The most important criticism against Schmidt's research has been that the randomization tests of his REG were either inadequate or inadequately described, in particular that they might not preclude short-term biases. The fact that scoring covaried with changes of target and the adoption of better control tests in later experiments reduce the force of this criticism.

The fact that Schmidt has been more consistently successful than other investigators in obtaining significant results in REG experiments, coupled with evidence from his research and others' that motivation in the absence of directed effort is sufficient to obtain such effects, raises the possibility that Schmidt himself is the source of the effects in his experiments.

The other major research program using REG methodology is being conducted under the supervision of Dr. Robert Jahn of Princeton University. So far, 22 subjects have completed 61 formal series involving a total of 569,450 runs of 200 trials each. Subjects were asked to use PK to either increase (PK+) or decrease (PK-) the generation of a binary target that was alternated from trial to trial. About one-third of the runs were baseline runs for which no PK influence was attempted.

Using the run as the unit of analysis, results indicated a significant displacement of the mean in the intended direction in both the PK+ and PK- conditions. Otherwise, the distributions were normal. The mean of the distribution of baseline runs did not differ significantly from MCE. Using the series as the unit, only the PK+ mean was significant. However, the

variance was significantly large for the PK- series and approached significance for the PK+ series. There was a suggestive restriction of variance in the baseline series.

The significant effects with the run as the unit were attributable to runs where the subject chose if the run was to be PK+, PK-, or baseline. It also seems to be the case that the results are largely if not exclusively attributable to one subject who contributed 14 of the 61 series. The results of this subject are independently significant, whereas the results of the remaining subjects generally are not. This subject also contributed the bulk of the significance in 34 exploratory series involving various minor changes in test parameters. It is possible that some of the other subjects in Jahn's research could have also achieved significant results had they completed as many runs as this one subject, but that remains to be demonstrated.

Internal analyses suggest that data selection by post-hoc classification of series as exploratory or optional stopping, if either had occurred, would not impact the positive conclusions from the research. Checks on proper functioning of the apparatus and failsafes against recording errors appear to have been extensive. On the other hand, no systematic efforts were made to monitor subject behavior during the test sessions. However, it would appear that data tampering would require computer sophistication on the part of the subject and knowledge of the system.

Finally, it is noted that it is not necessary to assume a causal influence on the functioning of the REG to explain Jahn's results. A simple alternative, which could in principle account for all the significant effects reported in the formal series (including the variance effects), is judicious "sorting" of a random ("chance") distribution of run scores into the PK+, PK-, and baseline categories. Whether this or a more straightforward "PK" model is preferable must be determined by further

research.

The Delmore Experiments

The Delmore experiments consisted primarily of restricted-choice card-guessing studies conducted with a male law student (B.D.) at J.B. Rhine's Institute for Parapsychology in the early 1970s. The principal investigators were Drs. B.K. Kanthamani and Edward Kelly.

The better controlled of the card-guessing methods was labeled "single-card clairvoyance" (SCC). The experimenter, who was seated at a desk, removed for each trial one of a large batch of ordinary playing cards from a drawer, placed it inside an opaque folder, and exposed the folder to B.D. who, seated on the other side of the desk, orally made his guess. Four formal series totaling 46 runs of 52 trials each were completed with this method.

The other method, called the shuffle method, involved the experimenter and B.D. each shuffling a deck of cards and then matching up the two sequences. Six formal series totaling 55 runs were completed using this method.

Results were evaluated by a statistic developed by Fisher which gave independent assessments of the scoring rates for number, suit, and the two combined (exact hits). For both methods, highly significant results were obtained for exact hits. B.D. was especially successful in predicting when he would achieve an exact hit: with the SCC method he made 20 such "confidence calls," of which 14 were exact hits; with the shuffle method, all 50 confidence calls were exact hits.

With the SCC method, it was found that on high-scoring runs B.D.'s misses revealed a systematic structure of errors or "confusions" that matched the structure of errors he made when briefly exposed tachistoscopically to slides of the playing cards, suggesting that the ESP and sensory stimuli underwent similar cognitive processing.

Finally, eight formal tests (5377 trials) using an REG of the Schmidt type also yielded highly significant results.

The major critic of the Delmore experiments has been statistician Persi Diaconis, who argued that the experiments should be discounted because they were not monitored by a magician. His concern was piqued because he had witnessed B.D. give an informal demonstration of alleged psychic abilities which convinced him that B.D. had utilized sleight-of-hand. Kelly responded that it was illegitimate to generalize from such an informal demonstration to formal experiments where sleight-of-hand could be precluded.

If B.D. were motivated to use magic tricks, it is conceivable that, in the case of the SCC method, he could have occasionally seen cards being transferred from the desk drawer to the opaque folder were he to have a concealed pocket mirror on his lap. However, an interview with Dr. Kanthamani suggests that the desk used in these experiments had a back, which would preclude this hypothesis. The shuffle method seems generally less secure against manipulation than the SCC method because in most cases the subject had physical contact with the call deck after knowing the target order. Sensory-cue or fraud hypotheses are strained in the case of the automated REG experiments, however. It is concluded that although the authors should have shown more sensitivity to the possible use of sleight-of-hand by B.D., Diaconis' critique lacks scientific weight in view of his failure to propose any counterhypotheses.

Target randomization was less than ideal with both card-guessing methods. However, the effects are too strong for artifacts of this type to be a likely explanation of the results. This conclusion is reinforced by empirical examination of the target sequences in the SCC series. Other sources of artifact appear to have been successfully precluded.

Correlational Studies

Although studies of the correlations between psi scores and psychological variables do not bear directly on the anomalous nature of the former, they can make important contributions by demonstrating the existence of a coherent class of events, identifying factors that may lead to improved reliability of psi scores, and serving as the building blocks for theories. However, due to the low reliability of psi scores, consistent confirmations of correlational findings cannot be expected. Thus, "real" effects can only be uncovered for those predictors which have been used frequently enough in psi research to provide a sufficiently large data base for meaningful meta-analysis.

The only predictors that have enjoyed sufficiently widespread use to spawn systematic meta-analyses have been the personality factors of extraversion and neuroticism, belief in psi (the "sheep-goat" variable), and hypnosis. In the case of extraversion, Sargent reported 19 of 54 relationships with ESP scores to be significant, with the extraverts scoring higher in 18 of the 19. Palmer found that 18 of 24 experimental series where subjects were tested individually or in pairs revealed more negative scores among the more neurotic subjects. A later survey of ten studies using the projective Defense Mechanism Test as the predictor yielded lower scoring by the more defensive subjects in all cases, with one-tailed significance achieved in seven of the ten. "Sheep" (believers) scored higher than "goats" in 18 of 21 series reviewed by Palmer. Schechter found that in 16 of 20 series subjects scored higher under hypnosis than in the "waking" state. All seven of the significant differences favored the hypnosis condition. In all these cases, the trends across studies were consistent to a statistically significant degree.

Evaluation focused on the validity, the interpretations, and the generality of the relationships. Procedural flaws which impacted on the integrity of the psi effects per se were considered to be of generally the

same type and prevalence as those uncovered by Hyman in his review of the ganzfeld experiments. Special consideration was given to a criticism raised by Charles Akers that subjects often knew their ESP scores before completing the psychological questionnaires. Although this flaw was rather widespread in the data base, further analyses suggested that it was unlikely to have accounted for the results.

Little research has been done to evaluate possible interpretations of the correlational effects reviewed in this chapter, although several have been suggested. The extraversion-ESP relationship has been attributed both to differences in cortical arousal and to adaptability to the social situation of psi testing. The hypnosis-ESP relationship has been attributed both to the implicit or explicit suggestions of success and to the altered state of consciousness induced by hypnotic induction procedures. Stanford noted various inadequacies in the control conditions of these studies that might bear on the interpretation of results, including lack of counterbalancing, nonrandom assignment of subjects to conditions, and experimenters not being blind to the treatment. Although these artifacts are considered less problematic than Stanford suggests, they cannot be completely discounted.

In the absence of a planned series of attempted replications, the robustness or generality of the relationships considered in this chapter cannot be assessed with great confidence, although the failure of one investigator (Thalbourne) to confirm the extraversion-ESP effect in a series of experiments argues for caution. Because the meta-analyses reviewed in this chapter generally compared the ratio of positive to negative outcomes rather than the proportion of significant outcomes per se, bias due to nonpublication of chance studies is considered unlikely. Of greater importance is the fact that the significant studies are not evenly distributed among the investigators who conducted them, although studies by the same investigator have been treated as independent in the meta-analyses.

Moreover, parapsychologists as a group constitute a rather specialized population when compared to scientists generally or even psychologists. The explanation of the "experimenter effect" in parapsychology is not yet known, but if adequate replicability is to be achieved, priority must be given to its elucidation.

Psi-Mediated Instrumental Response

One of the more serious attempts at theorizing in parapsychology is Rex Stanford's model of psi-mediated instrumental response (PMIR). The primary postulate of the model is that "...[an] individual, through extrasensory means, actively scans his environment for objects and events...which are relevant to his needs and that when such information is discovered he tends to respond to it in accordance with his typical dispositions toward such objects and events." (Stanford & Stio, 1976, p. 55) An important implication of the model is that psi can occur without the subject being aware of its occurrence or intending to use it. PMIR occurs as economically as possible through a variety of mediating vehicles, such as modification of the timing of an already selected response. The disposition toward PMIR is governed by the strength of the organism's needs, thereby linking the model to reinforcement theory in psychology. The model also applies to PK, which "can occur as a response to extrasensory or sensory information which has never been in the conscious focus of the PK...agent..." (Stanford, 1974b, p. 350)

Stanford published five related experiments to test the model. A covert ESP test was developed for these experiments. Subjects, generally males, were given a 10-item word-association test and the response latencies were measured. If the fastest (or, in some cases, slowest) latency was associated with a randomly selected key word, the subject would subsequently engage in a pleasant task; otherwise, he would engage in an unpleasant task. Thus, ESP could be used to "escape" the latter. The ESP score was

related to the standardized difference between the response latency to the key word and the mean response latency to all the words. In the one PK experiment, the method was simply to leave an REG running while the subject was performing the unpleasant task. If the REG produced a significant outcome, the subject escaped to the pleasant task.

Only in the PK experiment was the overall scoring level unequivocally significant. However, the primary objective of all the experiments was to test predictions from the PMIR model by manipulation of independent variables. The predictions fell into four classes: (1) PMIR scores will correlate with scores on standard (overt) psi tasks; (2) PMIR will most likely be facilitated by readily available cognitions (operationally defined as primary responses on the word-association test); (3) PMIR will be related to the strength of the need which it subserves; and (4) the direction of PMIR will be influenced by whether the subject has a positive or negative self-concept. All six tests of these predictions yielded results in the predicted direction, and three were significant.

The PMIR model and research program have not been addressed by outside critics. However, Stanford himself eventually abandoned the model because he found its "psychobiological" or cybernetic assumptions to be untenable. He replaced it with a simpler "conformance model" which did not assume any communication of information across a channel or assume that PK is guided in a step-by-step fashion by unconscious ESP.

The above problems notwithstanding, the PMIR model and the basic test paradigm seem sound. The major methodological objections were failure to check on the efficacy of the experimental manipulations and failure to report all potential tests of the hypotheses across experiments. How subjects were assigned to conditions was not fully specified in all cases. Finally, it would have been better to modify the PMIR model than to abandon it wholesale. Many "psychobiological" assumptions must be retained even in the conformance model. The later model suffers from ambiguity and the

research to which it has led does not follow from it uniquely.

Metal Bending

A renewed interest has recently developed among parapsychologists in large-scale (macro) PK effects, particularly metal bending. The most extensive research on metal bending has been conducted by Dr. John Hasted, who has worked with 20 subjects, mostly adolescents. Subjects were asked to bend or otherwise deform metal specimens (latchkeys or bars of aluminum alloy) without touching them. The specimens were attached to resistive strain gauges or (in later work) piezoelectric sensors. Signals from these devices were then amplified and registered on chart recorders.

Although actual bending was observed to have occurred in only a minority of sessions, anomalous signals with rapid rise times frequently appeared on the chart records. In some sessions, signals frequently appeared simultaneously from sensors separated up to several feet from each other. The fact that such "synchronous" signals seemed especially prevalent when the specimens were aligned vertically led Hasted to postulate a "surface of action" extending outward from the subject in a vertical plane.

Attempts to gain further information about the nature of the forces involved by locating multiple strain gauges across the width or along the length of a specimen generally failed to produce results consistent with simple extension, contraction, or bending hypotheses. This led Hasted to refer to his results more generally as "metal churning" as opposed to "metal bending." However, the effects seemed to be somewhat localized in the middle of specimens when strain gauges were distributed along their surfaces.

At least one subject seemed able to trigger electronic touch detectors without actually touching the specimen, suggesting that some of the ostensible strain effects may have been electrical in nature. Some unknown form of conduction of electrical charge from the subject's body through the atmosphere to the sensor was hypothesized on the basis of subsequent

research with this subject.

Physical contact with the specimens during the trials generally seems unlikely as an explanation of either the macro- or micro-effects, although explanations of the procedures were inadequate to preclude substitution of bent for unbent specimens. Localized artifactual influences on individual specimens (e.g., air currents from blowing) seem unlikely but were not thoroughly ruled out in all cases. They are not precluded by the claimed synchronicity of signals from different sensors, because the operational definition of synchronicity was not sufficiently precise to rule out radiation of an electromagnetic signal from one point to several sensors. Likewise, if one can trust Hasted's claim that the subject had no opportunity to interact directly with the chart recorder, the employment of dummy loads along with electrical shielding of the test channels minimize, although they do not rule out, more global artifacts.

Even if one grants the paranormal origins of the signals, Hasted's methodology makes it difficult to draw valid conclusions about their nature, including whether or not they truly represent strain. Use of an inadequately fast chart recorder, failure to adopt proper principles of experimental design, and failure to use statistical analyses are the most serious problems. In particular, it is impossible to distinguish basic physical characteristics of the phenomena from those correlated with preferences, attitudes, etc., of the subject or experimenter.

Although "no-touch" protocols are generally considered necessary in metal-bending research, reports by the French metallurgists Charles Crussard and J. Bouvaist of effects produced with touch by the subject Jean-Paul Girard are nonetheless worthy of scientific attention, in part because anomalous structural changes in the specimens were claimed. The authors described eight of the 20 trials conducted with Girard which they felt were conducted under adequately controlled conditions. The specimens were two bars of aluminum alloys, two stainless steel cylinders, and four Duralumin

plates. During the trials, Girard generally was allowed to touch and hold the specimens sometimes inside and sometimes outside a sealed glass tube, while at all times being observed by the experimenters.

Gross physical deformations (bending) were observed in only four of the specimens. Structural changes inconsistent both with a hypothesis of physical bending and with results from physically bent control samples were found for the stainless steel cylinder (excessive and anomalous distribution of magnetic martensite converted from austenite) and the Duralumin plates (a high density of small dislocation loops). The latter effect was found to be reproducible by a combination of shot-peening and polishing, however. More information about the base rates of anomalous results from the kinds of control tests the authors used would have been desirable.

The fact that Girard is known to possess conjuring skills demands caution in interpreting the above results. Despite extensive precautions by the authors, including consulations with magicians, video recording of trials, and the marking of test specimens, only in the case of the bending of one of the aluminum bars do the controls as reported seem to completely rule out the possibility of Girard substituting previously deformed specimens for the test specimens. Nonetheless, the assumptions that must be made to explain away these results seem rather farfetched, at least those assumptions of which this reviewer is aware.

Conclusions and Recommendations

My greatest difficulty in reviewing the research reports was the inadequacy of the methodological descriptions in most of them. It is likely that many of the criticisms raised by myself and others would have been answered if the reports had been more thorough. My general impression is that investigators trained as experimental psychologists gave better reports than those trained in the physical sciences or psychiatry. Perhaps the level of reporting adopted by the latter investigators is typical of

reporting in their own fields. Because of my background as a psychologist, I am more accustomed to the level of reporting found in the better psychology journals, and by that yardstick I found many of the reports I reviewed wanting. It seems to me that because so little is known about the effects studied in parapsychology, and because the effects are so often weak, unstable, and subject to "artifactual" influences, a higher degree of specificity is required in parapsychology than in more established disciplines. The Parapsychological Association is currently taking steps to improve the quality of reporting in the major journals.

However, I would not want to go to the other extreme and suggest that the documentation is so poor that the reports cannot be taken seriously. In virtually all cases, it was good enough to make critical reviews of the reports possible and potentially enlightening, and further methodological details sometimes came to light in exchanges with critics.

What are the conclusions that can be reached considering the research programs as a whole? With the possible exception of the PMIR and correlational programs, I think it fair to conclude that the results cannot be attributed to "chance." In most of the programs the cumulative results reach high levels of statistical significance. Because of these high levels, artifacts due to such things as violation of statistical independence, optional stopping, etc., are likely to be trivial and in most cases were shown to be trivial. Likewise, absurdly large numbers of relevant nonsignificant studies must be assumed in order to cancel out these trends. In short, something is clearly going on in these research projects that cannot be explained as statistical errors of measurement.

The question then becomes whether the results of these research programs can confidently be attributed to conventional mechanisms. At the risk of becoming a bore, I must again stress that the issue is not whether such mechanisms are <u>possible</u> but whether they are scientifically <u>adequate</u> as explanations. Unfortunately, I was only rarely able to appeal to internal

evidence in the data to make this assessment, and even less frequently was I able to cite relevant empirical evidence from other studies. Thus, to a large extent I had to resort to plausibility, a disturbingly subjective criterion.

The most prevalent of these conventional explanations as applied to the projects considered in this review involve inadequate or inadequately described randomization procedures. Problems include (1) crude or improper methods of target selection in the Delmore experiments and in some of the ganzfeld, remote viewing, and dream experiments; (2) inadequate randomization of judging materials in some of the remote viewing and ganzfeld experiments; and (3) baseline tests in some of Schmidt's REG experiments which did not adequately duplicate the procedures used in the experimental conditions.

However, it is doubtful that slight departures from randomness can adequately account for the magnitude and consistency of results in these research programs. Shuffling methods, for example, if undertaken with the care one would expect from a conscientious researcher, should be expected to yield sufficiently adequate randomization of targets for the purposes required. On the other hand, only in the Delmore SCC experiments was the adequacy of the actual target sequences evaluated. Research on the degree to which biased sequences actually effect results in standard psi testing paradigms would be useful. In the absence of such data, inadequate randomization represents a possible but not particularly plausible counterhypothesis for the results considered in this review.

The second major class of counterhypotheses are those concerning sensory cues (in ESP) or physical manipulation (in PK). These "artifacts" can be either incidental or intentional, i.e., fraudulent. If the latter, the fraud can be on the part of the subject(s) or investigator(s), or both. Opportunities for incidental sensory cueing seem possible in some of the ganzfeld and remote viewing experiments, but again they assume unusual

sensitivity on the part of subjects and judges that seems implausible and lacks an empirical basis. Where correlations between the possibility of such leakage and psi scores have been computed (as in the ganzfeld and hypnosis experiments), they have been found to be nonsignificant.

The concern about subject fraud is most acute in cases where the subjects are psychics with a reputation to defend (or promote) and/or are known to have conjuring skills. Of the research projects described here, the Delmore and macro-PK projects are the ones where these conditions seem most applicable. However, in neither case has anyone yet suggested a cheating mechanism that would be allowed by the methods as described in the reports.

Two other conventional possibilities, fraud on the part of an experimenter and some unknown artifact that cannot be inferred from the reports, cannot be assessed. However, the unreliability of ostensible psychic events (OPEs), and in particular the "experimenter effect," are likely to be reinforcing to anyone who is inclined to entertain these kinds of hypotheses. On the other hand, any interpretations of these frustrating characteristics of parapsychological data must be considered speculative at this time.

The bottom line is that the data reviewed in this report constitute genuine scientific anomalies for which no one has an adequate explanation or set of explanations. They are of scientific interest because, when taken at face value, they go beyond our present understanding of the most fundamental principles on which conventional science is based. In other words, if they are what they appear to be, their theoretical (and, eventually, their practical) implications are enormous. On the other hand, if the anomalies have conventional explanations, it is important to know this as well. Research in parapsychology has much in common with research in other scientific fields, particularly psychology. An understanding of "artifacts" in parapsychology could teach us much about how comparable artifacts impact

these other fields, where we might ordinarily be less inclined to look for them.

Progress in understanding OPEs has been frustratingly slow. There are many reasons for this, perhaps the most important of which is the elusiveness of OPEs. Yet little attention has been focused directly on solving this problem, which constitutes another important reason why progress in parapsychology has been so slow. Both parapsychologists and their critics have been preoccupied with determining whether there exist demonstrations of "psi" that preclude all conventional alternatives. As I have attempted to argue in Chapter 1, this approach is both futile and regressive. If progress is to be made, investigators must learn to accept the anomalies as such and then proceed to do process-oriented research to uncover the mechanisms (paranormal or conventional) that account for them. Research is especially needed to help us understand why OPEs occur so erratically and more frequently in some labs than in others. Although, as I mentioned in Chapter 1, much of the research in parapsychology is already process-oriented, this research is unsystematic and relatively underfunded. Although the projects covered in this report sometimes contained process-oriented elements, only in the PMIR and Delmore experiments were proper experimental designs consistently utilized. Funding agencies could provide a valuable service toward advancing knowledge in this area by discouraging investigations that merely give us one more demonstration of an anomaly and by encouraging investigations oriented toward uncovering the mechanisms that might be responsible for these anomalies or toward increasing their reliability.

Investigators most likely to make progress in this area are those who combine (1) knowledge of the literature in parapsychology and directly related subdisciplines in other fields, (2) experience with psi testing, (3) basic scientific competence, and (4) a sober, level-headed attitude. Researchers who are primarily propagandists for some particular

interpretation of the anomalies should be discouraged. A track record of getting the anomalies to occur in one's lab is obviously desirable but not necessary. A conventional theorist who cannot get the anomalies to occur could still make important research contributions by doing basic research on the conventional processes he or she thinks are responsible for others getting them. Likewise, if an unsuccessful paranormal theorist were to uncover a procedure that suddenly turned failure into success, the result could be especially important.

A Note on Applications

If the effects reviewed in this report represent paranormal human abilities, the potential for application is both obvious and significant. What is at issue is a qualitative increase in the capacity of man to interact with his environment, both in terms of acquiring information about it (ESP) and manipulating it (PK). In a sense, it is misleading to restrict discussion to particular applications such as medicine or police work. The fact is that these alleged abilities are applicable to the entire range of human activity. Society would never be the same if these abilities are real and could be brought under control.

Determining whether "psi" can be applied effectively in practical situations is the primary function of applied research in parapsychology. This is in contrast to basic research in parapsychology, where the objective is to understand the mechanism behind these effects and, in particular, whether or not the mechanism is "paranormal." These different objectives require different fundamental research strategies. Basic research requires controls against "normal" mechanisms, attempts to uncover correlates of the effects, and tests of hypotheses derived from paranormal theories. Strictly speaking, none of the above is required for applied research. In concrete terms, the purpose of applied research is to show whether psychics, doing whatever they do, are more successful at solving a particular problem than

the best currently available expertise or technology. Whether the psychic might be aided in achieving this objective by sensory cues, for example, is, from the practical point of view, irrelevant. (Of course, if it were exclusively sensory cues, the kinds of lofty outcomes alluded to in the first paragraph would not be expected.) For the above reasons, I prefer the more theoretically neutral term "applied intuition" over "applied psi" to label the process under study in applied parapsychological research.

One methodological implication of the preceding analysis is that some of the constraints placed upon psychics in basic research might be relaxed in applied research. Insofar as this improves the mood, confidence, and relaxation of the psychic, and insofar as these positive psychological attributes do indeed facilitate performance on "psi" tasks, psychics should be somewhat more successful in applied research contexts than in many basic research contexts. At the same time, it should be stressed that in other respects the methodology of applied research must be as rigorous as that demanded of basic research. In particular, the control expertise or technology must be rigorously defined and executed, the results from the psychic and control attempts must be directly and precisely comparable, and proper principles of experimental design must be employed.

A useful principle that can be, and has been, exploited in applied parapsychological research is the majority vote principle. It has been demonstrated on more than one occasion that a single subject who is acquiring a small degree of information can increase his or her reliability by repeatedly guessing the same target (e.g., Ryzl, 1966; Puthoff, in press). A related technique is to have multiple subjects attempt to acquire information about a single target. In applied research, this translates into having a group of psychics independently attempt to gather impressions about a single target or target event. Standard statistical techniques can be used to assess whether the degree of agreement among the psychics exceeds what would be expected by chance. If it does, then in theory at least, the

particular points on which the psychics agree should be the most accurate ones. An important implication of the majority-vote principle is that the small magnitude of the effects found in most psi experiments need not preclude their successful application. It is the poor reliability of the effects that creates the problem.

I have been able to find no applied research projects of sufficient extent and scientific merit to justify detailed review by the criteria set forth in Chapter 1, even granted the modifications presented above. The best research I could find were two studies by Martin Reiser who, in cooperation with the Los Angeles Police Department, attempted to assess the ability of local psychics to provide practically useful information about crimes (Reiser & Klyver, 1982; Reiser, Ludwig, Saxe, & Wagner, 1979). In each study, twelve psychics were given a piece of physical evidence in a concealed envelope from each of four crimes and asked to give their associations. (In the first study, the psychics were also allowed to see the evidence unconcealed.) In the second study, two control groups (college students and homicide detectives) were used. A double-blind protocol was applied in both experiments. Although the methods of analysis were crude, in neither experiment did the psychics succeed in providing useful information. (In the first study, a post-hoc analysis I conducted revealed that the four amateur psychics obtained significantly higher scores [low as they were] than did the eight professional psychics!)

A much more ambitious project was recently undertaken by Stephen Schwartz (1982), who asked a team of psychics, including Hella Hammid (see Chapter 3), to use remote viewing in an effort to locate buried architectural sites of ancient Egypt. These data were compared to judgments rendered by archaeological experts. It is difficult to be critical of this research because of the overwhelming logistical problems which the authors faced and which are discussed at great length in the book. Nonetheless, the remote viewings were not dramatically successful and the data were not

collected or evaluated in a way that allowed definitive conclusions to be drawn. This difficulty is exacerbated by the fact that the book was written more like a novel than a scientific report. However, the general approach which Schwartz is taking provides a good model for applied parapsychological research.

Perhaps the most important potential application of psi in the eyes of the general public is unorthodox healing. Incredibly, I am not aware of a single published report of a properly controlled experiment of psychic healing of humans. I am, however, aware of such a study recently completed at the Universtiy of Utrecht (The Netherlands) and which should soon be available.

Despite countless testimonials in the popular media to dramatically successful examples of applied intuition, the scientific evidence does not suggest that this intuition is sufficiently reliable to compete with or even usefully supplement presently available alternative methods. However, so little properly controlled evaluation research has been done that any firm conclusions about the efficacy of applied intuition appear premature. Although such evaluation research can and should be undertaken, it is my own opinion that applied intuition will never have a significant social impact until the mechanisms underlying this intuition (or psi) are better understood. Although such understanding is not logically necessary for successful application, and there are precedents for successful application in the absence of understanding, it seems to me that if such were the case in parapsychology successful application would already have become evident in the culture and routinely employed. Thus, it is my view that basic research should be given precedence over applied research in parapsychology at the present time.

REFERENCES

Akers, C. (1984). Methodological criticisms of parapsychology. In S. Krippner (Ed.), Advances in Parapsychological Research. Vol. 4 (pp. 112-164). Jefferson, NC: McFarland.

Alcock, J.E. (1981). Parapsychology: Science or Magic? Elmsford, NY: Pergamon.

Beloff, J. (1979). In defence of the "psychobiological" paradigm. In W.G. Roll (Ed.), Research in Parapsychology 1978 (pp. 11-12). Metuchen, NJ: Scarecrow.

Belvedere, E., & Foulkes, D. (1971). Telepathy and dreams: A failure to replicate. Perceptual and Motor Skills, 33, 783-789.

Bertini, M., Lewis, H.B., & Witkin, H.A. (1969). Some preliminary observations with an experimental procedure for the study of hypnagogic and related phenomena. In C.T. Tart (Ed.), Altered States of Consciousness, (pp. 95-114). Garden City, NY: Doubleday.

Bisaha, J., & Dunne, B.J. (1977). Precognitive remote viewing in the Chicago area: A replication of the Stanford experiment. In J.D. Morris, W.G. Roll, & R.L. Morris (Eds.), Research in Parapsychology 1976 (pp. 84-86). Metuchen, NJ: Scarecrow.

Bisaha, J.P., & Dunne, B.J. (1979). Multiple subject and long-distance precognitive remote viewing of geographical locations. In C.T. Tart, H.E. Puthoff, & R. Targ (Eds.), Mind At Large (pp. 109-124), New York: Praeger.

Blackmore, S. (1980). The extent of selective reporting of ESP ganzfeld studies. European Journal of Parapsychology, 3, 213-219.

Braud, W.G. (1978). Allobiofeedback: Immediate feedback for a psychokinetic influence upon another person's physiology. In W.G. Roll (Ed.), Research in Parapsychology 1977 (pp. 123-134). Metuchen, NJ: Scarecrow.

Braud, W.G. (1980). Lability and inertia in conformance behavior. Journal of the American Society for Psychical Research, 74, 297-318.

Braud, W.G., & Wood, R. (1977). The influence of immediate feedback on free-response GESP performance during ganzfeld stimulation. Journal of the American Society for Psychical Research, 71, 409-427.

Braud, W.G., Wood, R., & Braud, L.W. (1975). Free-response GESP performance during an experimental hypnagogic state induced by visual and acoustic ganzfeld techniques: A replication and extension. Journal of the American Society for Psychical Research, 69, 105-113.

Broad, C.D. (1953). Religion, Philosophy and Psychical Research. New York: Harcourt Brace.

Carpenter, J.C. (1971). The differential effect and hidden target differences consisting of erotic and neutral stimuli. Journal of the American Society for Psychical Research, 65, 204-214.

Casler, L. (1962). The improvement of clairvoyance scores by means of hypnotic suggestion. Journal of Parapsychology, 26, 77-87.

Cattell, R.B. (1965). The Scientific Analysis of Personality. Baltimore: Penguin.

Child, I.L. (in press). Psychology and anomalous observations: The question of ESP in dreams. American Psychologist.

Child, I.L., & Levi, A. (1979). Psi-missing in free-response settings. Journal of the American Society for Psychical Research, 73, 273-289.

Collins, H.M., & Pinch, T.J. (1982). Frames of Meaning: The Social Construction of Extraordinary Science. London: Rutledge & Kegan Paul.

Crussard, C., & Bouvaist, J. (1978). Expériences psychocinétiques sur éprouvettes métalliques. [Psychokinetic experiments with metal test samples]. Memoires scientifiques de la revue de méttalurgie. 13-23.

Diaconis, P. (1978). Statistical problems in ESP research. Science, 201, 131-136.

Diaconis, P. (1979). Rejoinder to Edward F. Kelly. Zetetic Scholar, No. 5, 29-31.

Diaconis, P. (1980). Replies to Edward F. Kelly and Charles T. Tart. Zetetic Scholar, No. 6, 131-132.

Dingwall, E.J. (1968). Abnormal Hypnotic Phenomena. London: Churchill (4 vols.).

Dunne, B.J., & Bisaha, J.P. (1979). Precognitive remote viewing in the Chicago area: A replication of the Stanford experiment. Journal of Parapsychology, 43, 17-30.

Dunne, B.J., Jahn, R.G., & Nelson, R.D. (1983). Precognitive remote perception. Technical Note PEAR 83003. Engineering Anomalies Research Laboratory, School of Engineering/Applied Science, Princeton University, Princeton, NJ. (178 pp.).

Epstein, R.A. (1977). The Theory of Gambling and Statistical Logic. New York: Academic Press.

Eysenck, H.J. (1960). The Structure of Human Personality (2nd Ed.). London: Methuen.

Eysenck, H.J. (1967). Personality and extra-sensory perception. Journal of the Society for Psychical Research, 44, 55-71.

Fisher, R.A. (1924). A method of scoring coincidences in tests with playing cards. Proceedings of the Society for Psychical Research, 34 (Part 91), 181-185.

Fisk, G.W., & West, D.J. (1958). Dice-casting experiments with a single subject. Journal of the Society for Psychical Research, 39, 277-287.

Foulkes, D., Belvedere, E., Masters, R.E.L., Houston, J., Krippner, S., Honorton, C., & Ullman, M. (1972). Long-distance, "sensory-bombardment" ESP in dreams: A failure to replicate. Perceptual and Motor Skills, 35, 731-734.

Glass, G.V., McGaw, B., & Smith, M.L. (1981). Meta-analysis in Social Research. Beverly Hills, CA: Sage Publications.

Grad, B., Cadoret, R.J., & Paul, G.I. (1961). The influence of an unorthodox method of treatment on wound healing in mice. International Journal of

Parapsychology, 3 (No. 2), 5-24.

Habel, M.M. (1976). Varying auditory stimuli in the ganzfeld: The influence of sex and overcrowding on psi performance. In J.D. Morris, W.G. Roll, & R.L. Morris (Eds), Research in Parapsychology 1975 (pp. 181-184). Metuchen, NJ: Scarecrow.

Hansel, C.E.M. (1966). ESP: A Scientific Evaluation. New York: Scribners.

Hansel, C.E.M. (1980). ESP and Parapsychology: A Critical Re-evaluation. Buffalo: Prometheus.

Hasted, J.B. (1976). An experimental study of the validity of metal-bending phenomena. Journal of the Society for Psychical Research, 48, 365-383.

Hasted, J.B. (1977). Physical aspects of paranormal metal bending. Journal of the Society for Psychical Research, 49, 583-607.

Hasted, J.B. (1978). Merkmale paranormaler metallbiegephänomene. Zeitschrift für Parapsychologie und Grenzgebiete der Psychologie, 20, 173-181.

Hasted, J.B. (1981). The Metal-Benders. London: Routledge.

Hasted, J.B., & Robertson, D. (1979). The detail of paranormal metal bending. Journal of the Society for Psychical Research, 50, 9-20.

Hasted, J.B., & Robertson, D. (1980). Paranormal action on metal and its surroundings. Journal of the Society for Psychical Research, 50, 379-398.

Hasted, J.B., & Robertson, D. (1981). Paranormal electrical effects. Psychoenergetic Systems, 4, 159-187.

Hasted, J.B., Robertson, D., & Arathoon, P. (1983). PKMB research with piezoelectric sensors. In W.G. Roll, J. Beloff, & R.A. White (Eds.), Research in Parapsychology 1982 (pp. 39-42). Metuchen, NJ: Scarecrow.

Honorton, C. (1975). Objective determination of information rate in psi tasks with pictorial stimuli. Journal of the American Society for Psychical Research, 69, 353-359.

Honorton, C. (1976). Length of isolation and degree of arousal as probable factors influencing information retrieval in the ganzfeld. In J.D. Morris, W.G. Roll, & R.L. Morris (Eds.), Research in Parapsychology 1975

(pp. 184-186). Metuchen, NJ: Scarecrow.

Honorton, C. (1977). Psi and internal attention states. In B.B. Wolman (Ed.), Handbook of Parapsychology (pp. 435-472). New York: Van Nostrand Reinhold.

Honorton, C. (1978). Psi and internal attention states: Information retrieval in the ganzfeld. In B. Shapin & L. Coly (Eds.), Psi and States of Awareness (pp. 79-90). New York: Parapsychology Foundation.

Honorton, C. (1983). Response to Hyman's critique of psi ganzfeld studies. In W.G. Roll, J. Beloff, & R.A. White (Eds.), Research in Parapsychology 1982 (pp. 23-26). Metuchen, NJ: Scarecrow.

Honorton, C. (1985). Meta-analysis of psi ganzfeld research: A response to Hyman. Journal of Parapsychology, 49, 51-91.

Honorton, C., & Harper, S. (1974). Psi-mediated imagery and ideation in an experimental procedure for regulating perceptual input. Journal of the American Society for Psychical Research, 68, 156-168.

Honorton, C., & Krippner, S. (1969). Hypnosis and ESP performance: A review of the experimental literature. Journal of the American Society for Psychical Research, 63, 214-252.

Hyman, R. (1979). Psychics and scientists: A review of Mind Reach. The Humanist, 37 (3).

Hyman, R. (1981). Further comments on Schmidt's PK experiments. Skeptical Inquirer, 5 (No. 3), 34-40.

Hyman, R. (1983). Does the ganzfeld experiment answer the critics' objections? In W.G. Roll, J. Beloff, & R.A. White (Eds.), Research in Parapsychology 1982 (pp. 21-23). Metuchen, NJ: Scarecrow.

Hyman, R. (1985). The ganzfeld psi experiment: A critical appraisal. Journal of Parapsychology, 49, 3-49.

Isaacs, J.D. (1984). Some aspects of performance at a psychokinetic task. Unpublished doctoral dissertation. University of Aston at Birmingham, England.

Jahn, R.G., Dunne, B.J., & Jahn, E.G. (1980). Analytical judging procedure for remote perception experiments. Journal of Parapsychology, 44, 207-231.

Jahn, R.G., Nelson, R.D., & Dunne, B.J. (1985, August). Variance effects in REG series score distributions. Paper presented at the Parapsychological Association convention, Tufts University, Medford, MA.

Johnson, M. (1973). A new technique of testing ESP in a real-life, high-motivational context. Journal of Parapsychology, 37, 210-217.

Johnson, M., & Haraldsson, E. (1984). The defense mechanism test as a predictor of ESP scores. Journal of Parapsychology, 48, 185-200.

Kanthamani, H., & Kelly, E.F. (1974a). Awareness of success in an exceptional subject. Journal of Parapsychology, 38, 355-382.

Kanthamani, H., & Kelly, E.F. (1974b). Card experiments with a special subject: I. Single-card clairvoyance. Journal of Parapsychology, 38, 16-26.

Kanthamani, H., & Kelly, E.F. (1975). Card experiments with a special subject: II. The shuffle method. Journal of Parapsychology, 39, 206-221.

Kanthamani, B.K. [H.], & Rao, K.R. (1973). Personality characteristics of ESP subjects: V. Neuroticism and ESP. Journal of Parapsychology, 37, 37-50.

Karnes, E.W., Ballou, J., Susman, E.P., & Swaroff, P. (1979). Remote viewing: Failures to replicate with control comparisons. Psychological Reports, 45, 963-973.

Karnes, E.W., & Susman, E.P. (1979). Remote viewing: A response bias interpretation. Psychological Reports, 44, 471-479.

Karnes, E.W., Susman, E.P., Klusman, P., & Turcotte, L. (1980). Failures to replicate remote-viewing using psychic subjects. Zetetic Scholar, No. 6, 66-76.

Kelly, E.F. (1979). Reply to Persi Diaconis. Zetetic Scholar, No. 5, 20-28.

Kelly, E.F. (1980). Response to Persi Diaconis's reply. Zetetic Scholar, No. 6, 121-127.

Kelly, E.F., & Kanthamani, B.K. [H.] (1972). A subject's efforts toward voluntary control. Journal of Parapsychology, 36, 185-197.

Kelly, E.F., Kanthamani, H., Child, I.L., & Young, F.W. (1975). On the relation between visual and ESP confusion structures in an exceptional ESP subject. Journal of the American Society for Psychical Research, 69, 1-31.

Kennedy, J.E. (1979). Consistent missing: A type of information-error in ESP. Journal of Parapsychology, 43, 113-128.

Krippner, S. (Ed.) (1977). Advances in Parapsychological Research. Vol. 1: Psychokinesis. New York: Plenum.

Krippner, S. (Ed.) (1978). Advances in Parapsychological Research. Vol. 2: Extrasensory Perception. New York: Plenum.

Krippner, S. (Ed.) (1982). Advances in Parapsychological Research. Vol. 3. New York: Plenum.

Krippner, S. (Ed.) (1984). Advances in Parapsychological Research. Vol. 4. Jefferson, NC: McFarland.

Krippner, S., Honorton, C., & Ullman, M. (1972). A second precognitive dream study with Malcolm Bessent. Journal of the American Society for Psychical Research, 66, 269-279.

Krippner, S., Honorton, C., & Ullman, M. (1973). An experiment in dream telepathy with the "Grateful Dead." Journal of the American Society of Psychosomatic Dentistry and Medicine, 20, 9-17.

Krippner, S., Honorton, C., Ullman, M., Masters, R., & Houston, J. (1971). A long-distance "sensory bombardment" study of ESP in dreams. Journal of the American Society for Psychical Research, 65, 468-475.

Krippner, S., & Ullman, M. (1970). Telepathy and dreams: A controlled experiment with electroencephalogram-electro-oculogram monitoring. Journal of Nervous and Mental Disease, 151, 394-403.

Krippner, S., Ullman, M., & Honorton, C. (1971). A precognitive dream study with a single subject. Journal of the American Society for Psychical

Research, 65, 192-203.

Layton, B.D., & Turnbull, B. (1975). Belief, evaluation, and performance on an ESP task. Journal of Experimental Social Psychology, 11, 166-179.

Lovitts, B.E. (1981). The sheep-goat effect turned upside down. Journal of Parapsychology, 45, 293-309.

Marks, D. (1981a). On the review of The Psychology of the Psychic: A reply to Dr. Morris. [Correspondence] Journal of the American Society for Psychical Research, 75, 197-203.

Marks, D. (1981b). Sensory cues invalidate remote viewing experiments. Nature, 292, 177.

Marks, D. (1982). Remote viewing revisited. Skeptical Inquirer, 6(4), 18-29.

Marks, D., & Kammann, R. (1978). Information transmission in remote viewing experiments. Nature, 274, 680-681.

Marks, D., & Kammann, R. (1980). The Psychology of the Psychic. Buffalo: Prometheus.

May, E.C., Radin, D.I., Hubbard, G.S., Humphrey, B.S., & Utts, J.M. (1985, August). Psi experiments with random number generators: An informational model. Paper presented at the Parapsychological Association convention, Tufts University, Medford, MA.

Morris, R.L. (1972). An exact method for evaluating preferentially matched free-response material. Journal of the American Society for Psychical Research, 66, 401-407.

Morris, R.L. (1980). Some comments on the assessment of parapsychological studies: A review of The Psychology of the Psychic by David Marks and Richard Kammann. Journal of the American Society for Psychical Research, 74, 425-443.

Moss, T., Paulson, M.J., Chang, A.F., & Levitt, M. (1970). Hypnosis and ESP: A controlled experiment. American Journal of Clinical Hypnosis, 13, 46-56.

Mosteller, F., & Bush, R.R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), Handbook of Social Psychology, Vol. 1 (pp. 289-334). Cambridge, MA: Addison-Wesley.

Nash, C.B., & Nash, C.S. (1958). Checking success and the relationship of personality traits to ESP. Journal of the American Society for Psychical Research, 52, 98-107.

Nelson, R.D., Dunne, B.J., & Jahn, R.G. (1984). An REG experiment with large data base capability, III: Operator related anomalies. Technical Note PEAR 84003, Princeton Engineering Anomalies Research, School of Engineering/Applied Science, Princeton University. (159 pp.)

Nicol, J.F., & Humphrey, B.M. (1953). The exploration of ESP and human personality. Journal of the American Society for Psychical Research, 47, 133-178.

Palmer, J. (1971). Scoring in ESP tests as a function of belief in ESP. Part I: The sheep-goat effect. Journal of the American Society for Psychical Research, 65, 373-408.

Palmer, J. (1972). Scoring in ESP tests as a function of belief in ESP. Part II: Beyond the sheep-goat effect. Journal of the American Society for Psychical Research, 66, 1-26.

Palmer, J. (1977). Attitude and personality traits in experimental ESP research. In B.B. Wolman (Ed.), Handbook of Parapsychology (pp. 175-201). New York: Van Nostrand Reinhold.

Palmer, J. (1978). Extrasensory perception: Research findings. In S. Krippner (Ed.), Advances in Parapsychological Research. Vol. 2: Extrasensory Perception (pp. 59-243). New York: Plenum.

Palmer, J. (1982). Methodological objections to the case for psi: Are formal control conditions necessary for the demonstration of psi? Journal of Indian Psychology, 4, 13-18.

Palmer, J. (1983). Sensory contamination of free-response ESP targets: The greasy fingers hypothesis. Journal of the American Society for Psychical

Research, 77, 101-113.

Palmer, J. (1985, August). Terminological poverty in parapsychology: Two examples. Paper presented at the Parapsychological Association convention, Tufts University, Medford, MA.

Palmer, J., & Kramer, W. (1984). Internal state and temporal factors in psychokinesis. Journal of Parapsychology, 48, 1-25.

Palmer, J., & Kramer, W. (in press). Sensory identification of contaminated free-response ESP targets: Return of the greasy fingers. Journal of the American Society for Psychical Research.

Palmer, J., & Lieberman, R. (1975). The influence of psychological set on ESP and out-of-body experiences. Journal of the American Society for Psychical Research, 69, 193-213.

Parker, A. (1975). Some findings relevant to the change in state hypothesis. In J.D. Morris, W.G. Roll, & R.L. Morris (Eds.), Research in Parapsychology 1974 (pp. 40-42). Metuchen, NJ: Scarecrow.

Parker, A., & Wiklund, N. (1982). The ganzfeld: A methodological evaluation of the claims for a repeatable ESP experiment. Unpublished manuscript.

Puthoff, H.E. (in press). Computer assisted psi amplification. In D.H. Weiner & D.I. Radin (Eds.), Research in Parapsychology 1985. Metuchen, NJ: Scarecrow.

Puthoff, H.E., & Targ, R. (1979). A perceptual channel for information transfer over kilometer distances: Historical perspective and recent research. In C.T. Tart, H.E. Puthoff, & R. Targ (Eds.), Mind at Large, (pp.13-76). New York: Praeger.

Puthoff, H.E., & Targ, R. (1981). Rebuttal of criticisms of remote viewing experiments. Nature, 292, 388.

Puthoff, H.E., Targ, R., & May, E. (1979, January). Experimental psi research: Implications for physics. Paper presented at the 145th National Meeting of the American Association for the Advancement of Science, Houston, TX.

Raburn, L. (1975). Expectation and transmission factors in psychic functioning. Unpublished honors thesis, Tulane University, New Orleans, LA.

Randi, J. (1975). The Truth about Uri Geller. Buffalo: Prometheus.

Randi, J. (1982). Flim-flam. Buffalo: Prometheus.

Rao, K.R. (1965). The bidirectionality of psi. Journal of Parapsychology, 29, 230-250.

Rauscher, E.A., & Rubik, B.A. (1980). Effects of motility behavior and growth rate of Salmonella typhimurium in the presence of a psychic subject. In W.G. Roll (Ed.), Research in Parapsychology 1979 (pp. 140-142).

Reiser, M., & Klyver, N. (1982). A comparison of psychics, detectives, and students in the investigations of major crimes. In M. Reiser (Ed.), Police Psychology: Collected Papers (pp. 260-267). Los Angeles: LAMI Publications.

Reiser, M., Ludwig, L., Saxe, S., & Wagner, C. (1979). An evaluation of the use of psychics in the investigation of major crimes. Journal of Police Science and Administration, 7, 18-25. Metuchen, NJ: Scarecrow.

Rhine, J.B. (1948). Conditions favoring success in psi tests. Journal of Parapsychology, 12, 58-75.

Rhine, J.B. (1974). Comments: A new case of experimenter unreliability. Journal of Parapsychology, 38, 215-225.

Rhine, J.B., & Pratt, J.G. (1957). Parapsychology: Frontier Science of the Mind. Springfield, IL: Charles Thomas.

Rhine, J.B., Pratt, J.G., Stuart, C.E., Smith, B.M., & Greenwood, J.A. (1940). Extra-sensory Perception After Sixty Years. Boston: Bruce Humphries.

Roll, W.G. (1966). ESP and memory. International Journal of Neuropsychiatry. 2, 505-521.

Rosenthal, R. (1966). Experimenter effects in behavioral research. New York: Appleton-Century-Crofts.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. Psychological Bulletin, 86, 638-641.

Rosenthal, R., & Rubin, D.B. (1984). Multiple contrasts and ordered Bonferroni procedures. Journal of Educational Psychology, 76, 1028-1034.

Ryzl, M. (1962). Training the psi faculty by hypnosis. Journal of the Society for Psychical Research, 41, 234-252.

Ryzl, M. (1966). A model of parapsychological communication. Journal of Parapsychology, 30, 18-30.

Sargent, C.L. (1978). Hypnosis as a psi-conducive state: A controlled replication study. Journal of Parapsychology, 42, 257-275.

Sargent, C.L. (1980). A note on personality differences between psi-conducive and psi-inhibitory ganzfeld experimenters. In W.G. Roll (Ed.), Research in Parapsychology 1979 (pp. 114-115). Metuchen, NJ: Scarecrow.

Sargent, C.L. (1981). Extraversion and performance in "extra-sensory perception" tasks. Personality and Individual Differences, 2, 137-143.

Saunders, D.R. (1985). On Hyman's factor analysis. Journal of Parapsychology, 49, 86-88.

Schechter, E.I. (1984). Hypnotic induction vs. control conditions: Illustrating an approach to the evaluation of replicability in parapsychological data. Journal of the American Society for Psychical Research, 78, 1-27.

Scheffé, H. (1959). The Analysis of Variance. New York: Wiley.

Schlitz, M., & Gruber, E. (1980). Transcontinental remote viewing. Journal of Parapsychology, 44, 305-317.

Schlitz, M., & Gruber, E. (1981). Transcontinental remote viewing: A rejudging. Journal of Parapsychology, 45, 233-237.

Schlitz, M., & Haight, J.M. (1984). Remote viewing revisited: An intrasubject replication. Journal of Parapsychology, 48, 39-49.

Schmeidler, G.R. (1973). PK effects upon continuously recorded temperature. Journal of the American Society for Psychical Research, 67, 325-340.

Schmeidler, G.R., & McConnell, R.A. (1973/1958). ESP and Personality Patterns. Westport, CT: Greenwood Press.

Schmidt, H. (1969a). Clairvoyance tests with a machine. Journal of Parapsychology, 33, 300-306.

Schmidt, H. (1969b). Precognition of a quantum process. Journal of Parapsychology, 33, 99-108.

Schmidt, H. (1970a). PK experiments with animals as subjects. Journal of Parapsychology, 34, 255-261.

Schmidt, H. (1970b). A PK test with electronic equipment. Journal of Parapsychology, 34, 175-181.

Schmidt, H. (1970c). A quantum mechanical random number generator for psi tests. Journal of Parapsychology, 34, 219-224.

Schmidt, H. (1973). PK tests with a high-speed random number generator. Journal of Parapsychology, 37, 105-118.

Schmidt, H. (1974). Comparison of PK action on two different random number generators. Journal of Parapsychology, 38, 47-55.

Schmidt, H. (1975). Toward a mathematical theory of psi. Journal of the American Society for Psychical Research, 69, 301-319.

Schmidt, H. (1976). PK effect on pre-recorded targets. Journal of the American Society for Psychical Research, 70, 267-291.

Schmidt, H. (1978). A take-home test in PK with pre-recorded targets. In W.G. Roll (Ed.), Research in Parapsychology 1977 (pp. 31-36). Metuchen, NJ: Scarecrow.

Schmidt, H. (1979a). Search for psi fluctuations in a PK test with cockroaches. In W.G. Roll (Ed.), Research in Parapsychology 1978 (pp. 77-78). Metuchen, NJ: Scarecrow.

Schmidt, H. (1979b). Use of stroboscopic light as rewarding feedback in a PK test with pre-recorded and momentarily-generated random events. In W.G. Roll (Ed.), Research in Parapsychology 1978 (pp. 115-117). Metuchen, NJ: Scarecrow.

Schmidt, H. (1981). PK tests with pre-recorded and pre-inspected seed numbers. Journal of Parapsychology, 45, 87-98.

Schmidt, H., Morris, R.L., & Rudolph, L. (in press). Channeling evidence for a psychokinetic effect to independent observers. Journal of Parapsychology.

Schmidt, H., & Pantas, L. (1972). Psi tests with internally different machines. Journal of Parapsychology, 36, 222-232.

Schwartz, S.A. (1983). The Alexandria Project. New York: Dell.

Scott, C. (1972). On the evaluation of verbal material in parapsychology: A discussion of Dr. Pratt's monograph. Journal of the Society for Psychical Research, 46, 79-90.

Scott, C. (1982). No "remote viewing." Nature, 298, 414.

Smith, M., Tremmel, L., & Honorton, C. (1976). A comparison of psi and weak sensory influences on ganzfeld mentation. In J.D. Morris, W.G. Roll, & R.L. Morris (Eds.), Research in Parapsychology 1975 (pp. 191-194). Metuchen, NJ: Scarecrow.

Smith, M.J. (1972). Paranormal effects on enzyme activity. Human Dimensions 1 (No. 2), 15-19.

Sondow, N. (1979). Effects of associations and feedback on psi in the ganzfeld: Is there more than meets the judge's eye? Journal of the American Society for Psychical Research, 73, 123-150.

Stanford, R.G. (1967). Response bias and the correctness of ESP test responses. Journal of Parapsychology, 31, 280-289.

Stanford, R.G. (1974a). An experimentally testable model for spontaneous psi events: I. Extrasensory events. Journal of the American Society for Psychical Research, 68, 34-57.

Stanford, R.G. (1974b). An experimentally testable model for spontaneous psi events: II. Psychokinetic events. Journal of the American Society for Psychical Research, 68, 321-356.

Stanford, R.G. (1977). Conceptual frameworks of contemporary psi research. In B.B. Wolman (Ed.), Handbook of Parapsychology (pp. 823-858). New York:

Van Nostrand Reinhold.

Stanford, R.G. (1978). Toward reinterpreting psi events. Journal of the American Society for Psychical Research, 72, 197-214.

Stanford, R.G. (1979). The influence of auditory ganzfeld characteristics upon free-response ESP performance. Journal of the American Society for Psychical Research, 73, 253-272.

Stanford, R.G. (1982). On matching the method to the problem: Word-association and signal-detection methods for the study of cognitive factors in ESP tasks. In B. Shapin & L. Coly (Eds.), Parapsychology and the Experimental Method (pp. 1-18). New York: Parapsychology Foundation.

Stanford, R.G. (in press). Altered internal states and parapsychological research: Retrospect and prospect. In D.H. Weiner & D.I. Radin (Eds.), Research in Parapsychology 1985. Metuchen, NJ: Scarecrow.

Stanford, R.G., & Associates (1976). A study of motivational arousal and self-concept in psi-mediated instrumental response. Journal of the American Society for Psychical Research, 70, 167-178.

Stanford, R.G., & Rust, P. (1977). Psi-mediated helping behavior: Experimental paradigm and initial results. In J.D. Morris, W.G. Roll, & R.L. Morris (Eds.), Research in Parapsychology 1976. (pp. 109-110). Metuchen, NJ: Scarecrow.

Stanford, R.G., & Stio, A. (1976). A study of associative mediation in psi-mediated instrumental response. Journal of the American Society for Psychical Research, 70, 55-64.

Stanford, R.G., & Thompson, G. (1974). Unconscious psi-mediated instrumental response and its relation to conscious ESP performance. In W.G. Roll, R.L. Morris, & J.D. Morris (Eds.), Research in Parapsychology 1973 (pp. 99-103). Metuchen, NJ: Scarecrow.

Stanford, R.G., Zenhausern, R., Taylor, A., & Dwyer, M.A. (1975). Psychokinesis as psi-mediated instrumental response. Journal of the American Society for Psychical Research, 69, 127-133.

Stokes, D.M. (1978). Review of Research in Parapsychology 1976. Journal of Parapsychology, 42, 70-76.

Stokes, D.M. (1982). Review of The Metal-Benders. Journal of the American Society for Psychical Research, 76, 59-67.

Taddonio, J.L. (1975). Attitudes and expectancies in ESP scoring. Journal of Parapsychology, 39, 289-296.

Targ, R., & Puthoff, H.E. (1977). Mind-reach. New York: Delacorte.

Targ, R., Puthoff, H.E., & May, E.C. (1979). Direct perception of remote geographical locations. In C.T. Tart, H.E. Puthoff, & R. Targ (Eds.), Mind At Large (pp. 78-106). New York: Praeger.

Tart, C.T. (1980). [Comments (on Karnes, et al., 1980)] Zetetic Scholar, No. 6, 85-86.

Tart, C.T., Puthoff, H.E., & Targ, R. (1980). Information transmission in remote viewing experiments. Nature, 284, 191.

Terry, J.C. (1976). Comparison of stimulus duration in sensory and psi conditions. In J.D. Morris, W.G. Roll, & R.L. Morris (Eds.), Research in Parapsychology 1975 (pp. 179-181). Metuchen, NJ: Scarecrow.

Terry, J., & Schmidt, H. (1978). Conscious and subconscious PK tests with pre-recorded targets. In W.G. Roll (Ed.), Research in Parapsychology 1977 (pp. 36-41). Metuchen, NJ: Scarecrow.

Terry, J., Tremmel, L., Kelly, M., Harper, S., & Barker, P.L. (1976). Psi information rate in guessing and receiver optimization. In J.D. Morris, W.G. Roll, & R.L. Morris (Eds.), Research in Parapsychology 1975 (pp. 194-198). Metuchen, NJ: Scarecrow.

Thalbourne, M.A. (1981). Extraversion and the sheep-goat variable: A conceptual replication. Journal of the American Society for Psychical Research, 75, 105-119.

Thalbourne, M.A. (1982). A Glossary of Terms Used in Parapsychology. North Pomfret, VT: William Heinemann.

Thalbourne, M.A., Beloff, J., & Delanoy, D. (1982). A test for the "extraverted sheep versus introverted goats" hypothesis. In W.G. Roll, R.L. Morris, & R.A. White (Eds.), Research in Parapsychology 1981 (pp. 155-156). Metuchen, NJ: Scarecrow.

Thalbourne, M.A., Beloff, J., Delanoy, D., & Jungkuntz, J.H. (1983). Some further tests of the extraverted sheep versus introverted goats hypothesis. In W.G. Roll, J. Beloff, & R.A. White (Eds.), Research in Parapsychology 1982 (pp. 199-200). Metuchen, NJ: Scarecrow.

Thalbourne, M.A., & Jungkuntz, J.H. (1983). Extraverted sheep versus introverted goats: Experiments VII and VIII. Journal of Parapsychology, 47, 49-51. (Abstract)

Tyrrell, G.N.M. (1936). Further research in extra-sensory perception. Proceedings of the Society for Psychical Research, 44 (Part 147), 99-168.

Ullman, M., & Krippner, S. (1968, August). Experimentally induced telepathic dreams with EEG-REM monitoring: The Van de Castle study. Paper presented at the Parapsychological Association convention, University of Freiburg, Freiburg, West Germany.

Ullman, M., & Krippner, S. (1970). Dream Studies and Telepathy: An Experimental Approach. Parapsychological Monographs No. 12. New York: Parapsychology Foundation.

Ullman, M., & Krippner, S. (1978). Experimental dream studies. In M. Ebon (Ed.), The Signet Handbook of Parapsychology (pp. 409-422). New York: New American Library.

Ullman, M., Krippner, S., & Feldstein, S. (1969). Experimentally-induced telepathic dreams: Two studies using EEG-REM monitoring techniques. In G.R. Schmeidler (Ed.), Extrasensory Perception (pp. 137-161). New York: Atherton.

Ullman, M., Krippner, S., & Vaughan, A. (1973). Dream Telepathy. New York: MacMillan.

Van de Castle, R.L. (1969). The facilitation of ESP scores through hypnosis. American Journal of Clinical Hypnosis, 12, 37-56.

Van de Castle, R.L. (1977). Sleep and dreams. In B.B. Wolman (Ed.), Handbook of Parapsychology (pp. 471-499), New York: Van Nostrand Reinhold.

Walker, E.H. (1975). Foundations of paraphysical and parapsychological phenomena. In L. Oteri (Ed.), Quantum Physics and Parapsychology (pp. 1-44). New York: Parapsychology Foundation.

Weiner, D.H., & Geller, J. (1984). Motivation as the universal container: Conceptual problems in parapsychology. Journal of Parapsychology, 48, 27-37.

Wolman, B.B. (1977). (Ed.) Handbook of Parapsychology. New York: Van Nostrand Reinhold.

Wood, R.H. (1982). On the importance of correct mechanics in paranormal research. Journal of the Society for Psychical Research, 51, 246-249.

Zusne, L., & Jones, W.H. (1982). Anomalistic Psychology: A Study of Extraordinary Phenomena of Behavior and Experience. Hillsdale, NJ: Lawrence Erlbaum.