SECRET

PAG-TA-1068-SL

DEFENSE
INTELLIGENCE
AGENCY

PROJECT STAR GATE
RESEARCH AND PEER REVIEW PLAN (U)

JUNE 1994

NOFORN
SECRET
LIMDIS

STAR GATE

PROJECT STAR GATE
RESEARCH AND PEER REVIEW PLAN (U)

This document was prepared by the
Technology Assessment and Support Activity
Office for Ground Forces
Directorate for Military Assessments
National Military Intelligence Production Center
Defense Intelligence Agency

Date of Publication
June 1994

**REPRODUCTION REQUIRES
APPROVAL OF ORIGINATOR
OR HIGHER DOD AUTHORITY**

LIMITED DISSEMINATION

**FUTHER DISSEMINATION
ONLY AS DIRECTED BY DIA/PAG
OR HIGHER DOD AUTHORITY**

CLASSIFIED BY MULTIPLE SOURCES
DECLASSIFY ON OADR

**SECRET
NOT RELEASABLE TO FOREIGN NATIONALS
STAR GATE**

OUTLINE

(U) EXECUTIVE SUMMARY:

(S/NF/SG/LIMDIS) In compliance with the Congressional conferees' request (Appendix A), DIA proposes to develop a multi-year research and development program, subject to rigorous scientific and technical oversight, to demonstrate the scientific validity of the STAR GATE program, and that results of military and intelligence value can be obtained in a cost-effective manner using anomalous mental phenomena (AMP).

(S/NF/SG/LIMDIS) This proposed program, if successfully implemented, will:

- Identify the underlying mechanisms of AMP.
- Establish the limits of operational usefulness of AMP.
- Determine the degree to which foreign activities in AMP represents a threat to national security.
- Lead to the development of countermeasures to neutralize this threat.
- Use research findings to improve operational activities.
- Develop data fusion criteria to integrate AMP results with other intelligence sources.

(S/NF/SG/LIMDIS) Due to the diversity of the STAR GATE mission/objectives, both external resources and in-house expertise are required. Since this Activity possesses no in-house R&D capability, an absolute need for external R&D support is required to meet Congressional concerns which are addressed in this program plan. A balance will be maintained between external and in-house activities, and every effort will be made to integrate and link these activities where appropriate. The external aspect permits a wide range of expertise covering many disciplines to be focused on this area; this also has the benefit of ensuring peer group review and of facilitating a variety of scientific interactions. In-house personnel with a wide-range of expertise in this phenemenology will need to be retained to make this proposed plan work.

(S/NF/SG/LIMDIS) In order to fulfill Congressional Direction, the DIA proposes to convene a Scientific Evaluation Panel (SEP) composed of representatives from each of the Service Scientific Advisory Boards. The purpose of the SEP is to review and validate the methodology outlined in the plan in order to address the cost-effectiveness and performance criteria for the

STAR GATE program's research and development objectives and to propose recommendations as to which objectives should be pursued and the program scope required to achieve those objectives. If the SEP determines that objectives in the plan are viable and executable, the General Defense Intelligence Program (GDIP) Manager will complete this initiative with others for limited available resources remaining in the program.

(U) The proposed ongoing R&D effort will be reviewed every two years by the SEP to determine whether the STAR GATE program can show results that are cost-effective and satisfy reasonable performance criteria.

(C) An annual report will document the current operational, technical and administrative status of the program.

I. (U) <u>INTRODUCTION</u>:

(S/NF/SG/LIMDIS) This program plan was developed in response to a Defense Authorization Conference, Congressionally Directed Action (CDA) to prepare a long-term systematic and comprehensive research and peer review plan in order to investigate anomalous mental phenomena (AMP), and to apply program research results to potential operational activities. This plan also describes key in-house activities along with an appropriately integrated basic and applied external research support effort.

(S/NF/SG/LIMDIS) Specifically, this program plan represents DIA's view on how best to proceed with both in-house activities and external research support for the period of FY95 through FY99. Research findings, both domestic and foreign, and results from operational activities may lead to updates of this plan in order to reflect improved phenomena understanding and to pursue follow-on research and/or application directions.

(S/NF/SG/LIMDIS) A underlying and fundamental premise governing the implementation of this program plan is that a well-integrated interdisciplinary approach is considered to be the most appropriate strategy for conducting research in this diverse field. Consequently, this plan includes a wide variety of research topics which are based on recent findings from leading-edge pursuits in other disciplines that are suspected of being germane for STAR GATE. Other topics are derived from a review of worldwide research, consultations with leading area experts, and on insights gained from previous research and application activities associated with the STAR GATE program.

(S/NF/SG/LIMDIS) This program plan also includes recommended proposed FY funding which will allow for the STAR GATE program to show results that are cost effective and will at the same time satisfy reasonable program performance criteria. The implementation of this program plan will preclude the reoccurrence of the yearly cyclical activity of project start-up, limited progress, followed by anticipated project shut-down which previously inhibited program activity.

(S/NF/SG/LIMDIS) In sum, the implementation of this research and peer review plan will allow DIA to successfully accomplish identified R&D activities which, in-turn, will enhance the capability of STAR GATE personnel to engage in operational activities and to assess the work done by potential adversaries, thereby, reducing the risk potential for a technological surprise.

(U) Terminology and definitions are discussed at Appendix B.

II. (U) <u>PLAN OBJECTIVES</u>:

(S/NF/SG/LIMDIS) The objective of this follow-on research and peer review plan is to further develop phenomena understanding and/or validation, in applications understanding, and in operational feasibility evaluation. This continued work will have a direct bearing on DIA's ability to both assess the significance of foreign research and to perform a systematic review of potential applications regarding this phenomena.

(S/NF/SG/LIMDIS) Accomplishment of the various activities identified in this plan will further enhance threat assessment of foreign achievements in this area, and will help achieve the potential for U.S. military/intelligence applications on select tasks as a supplement to HUMINT operations.

(U) It is anticipated that this plan will assist decision makers in their review and consideration of future directions for this field, and that this plan can begin formal implementation starting in FY95.

(S/NF/SG/LIMDIS) In compliance with the Congressional conferees' request, DIA recommends that a period of six to nine months be set aside at the beginning of this new program for the purpose of identifying the most promising and cost-effective experiments to be conducted under the program to meet the overall research objectives outlined below. It is further suggested that a series of small working groups consisting of scientific experts from a variety of pertinent disciplines meet during this time period to accomplish this end. Their suggestions will be presented to the STAR GATE Scientific Oversight Committee for final approval.

III. (U) <u>SIGNIFICANCE OF EFFORT</u>:

(S/NF/SG/LIMDIS) STAR GATE is a dynamic approach for pursuing the largely unexplored area of human consciousness and subconsciousness interaction. Its scope is comprehensive; a wide range of phenomenological issues are examined that include psychological, physiological/neurophysiological, physics and other leading-edge scientific areas. Although broad in scope, STAR GATE is well grounded due to its solid independent scientific review base. STAR GATE is based on a dynamic style in all its endeavors, especially in its pursuit of on-going foreign activities in this area.

(S/NF/SG/LIMDIS) One of the tasks previously levied on DIA by the FY91 Defense Authorization Act was to develop a long-range comprehensive plan for investigating parapsychological phenomena. This task was one of several objectives included in a new program for this phenomenological area that identified DIA as executive agent. Moreover the FY91 Defense Authorization Act authorized for DIA a funding level of $2 million for DIA in order to initiate this new program. As a result, a balanced and integrated plan to include operations, foreign assessment, and research and development was implemented . In addition, a new DIA limited dissemination (LIMDIS) program, codeword STAR GATE, was established in order to accomplish the objectives that were set forth in this plan.

SG1B

[redacted]

(S/NF/SG/LIMDIS) The external research support conducted under monies appropriated to date comes to a close in the March/April 1994 time-frame. The impact of this is that if research activities utilizing human subjects are interrupted, it has generally been necessary to begin again instead of later resuming activities from the point of termination. Consequently, it is important for the STAR GATE program to remain stable. Research involving human use differs considerably from that involving physical systems. For example, data from human subjects cannot be collected nor analyzed as rapidly, in that additional empirical data is often required to reach analytical conclusions. This type of data analysis utilizing human subjects can only be achieved with an in-place, uninterrupted, multi-year research and development program. Therefore, should it be decided to go forward with this program, it should be done in a timely fashion.

(S/NF) The funding allocation for external research received by STAR GATE in FY91 and continued through FY93 permitted several important research areas to be initiated and continued. It is anticipated that results of this research will assist in clarifying some of the possible future research directions; consequently, not all long-range research possibilities can identified in this plan. However, most all of the major investigation areas can be addressed, and many of the specifics can be identified with reasonable confidence. Figure 1 presents an overview of overall research objectives for both Anomalous Cognition (AC) and Anomalous Perturbation (AP) which will be considered for inclusion in this program.

(S/NF) Previous basic research activities from FY91 through FY93 focused on the following; (1) validating findings from previous magnetoencephalograph (MEG) research and initiating new work with a variety of conditions and individuals; (2) performing a variety of anomalous cognition (AC) experiments to determine potential correlations (e.g., target type, environmental factors); (3) developing various theoretical constructs that might be testable and that could help explain the phenomena; (4) examining effects of altered states on data quality; (5) initiating review of and research into the energetics area; and (6) examining various application possibilities (e.g., communication, search).

(U) Results from previous basic and applied research activity have been factored into this research and development plan and provide the basis upon which further R&D efforts will be built.

IV. (U) PLAN OVERVIEW:

A. (U) BASIC RESEARCH OBJECTIVES

(S/NF/SG/LIMDIS) The objective of basic research is to understand the fundamental, underlying mechanisms for AMP. To achieve this objective in an efficient way, basic research of the detection mechanism should begin in a conservative direction. That is, assume that a putative "sensorial" system exists for AMP and that it most likely will behave similarly to those common elements which are known through the five senses. This conservative approach generalizes to understand the source of AMP and its propagation mechanisms.

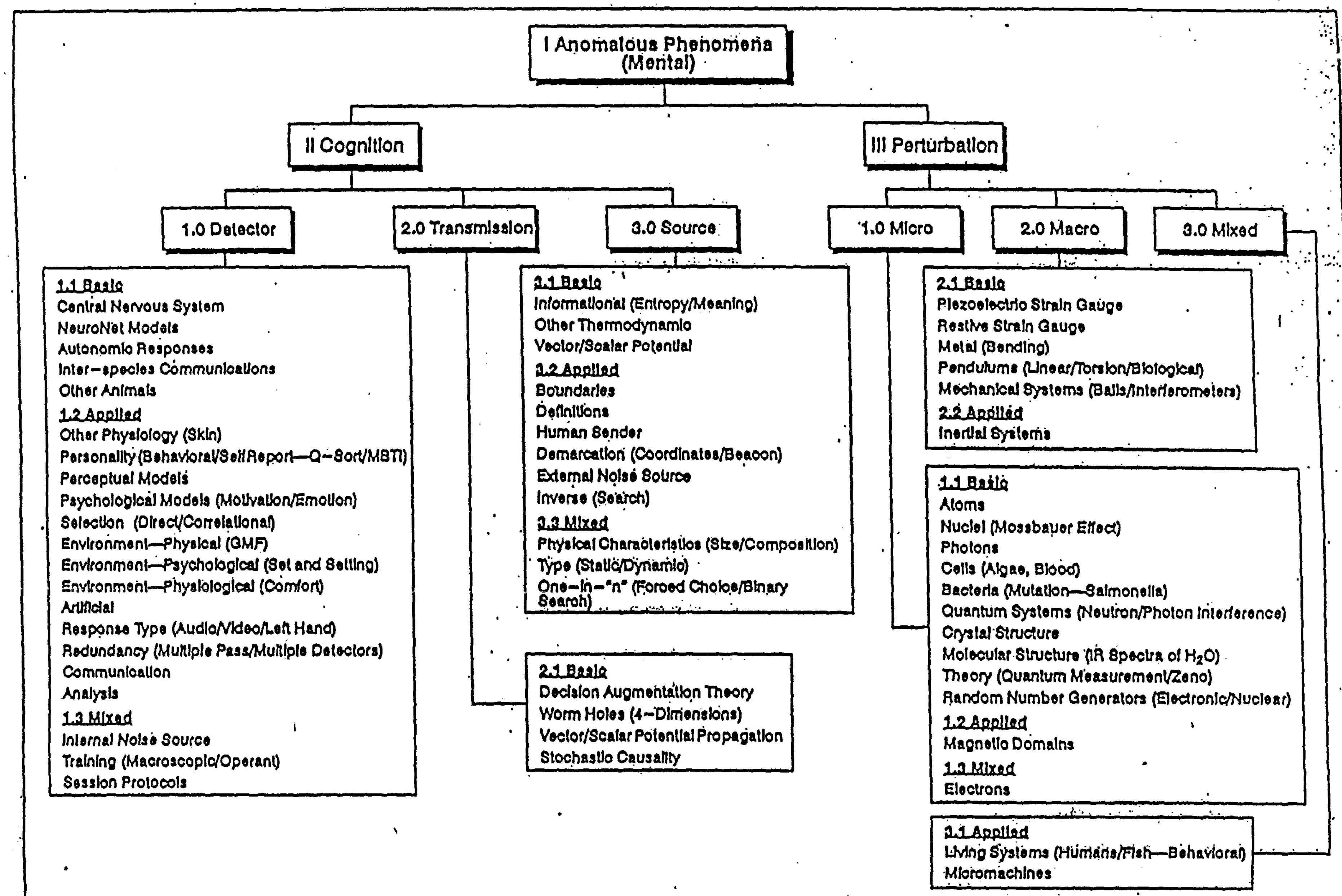

Figure 1 (U) Research Overview

B. (U) APPLIED RESEARCH OBJECTIVES

(S/NF/SG/LIMDIS) The objective of applied research is to improve AMP functioning to its maximum possible limit. To realize this objective, it is critical to define AMP output measures that are consistent with either a laboratory setting and/or an operational environment. The approach should also reflect scientific conservatism. In investigating any single variable (e.g., different training methodologies) all other variables should remain as constant as possible (e.g., use the same individuals and known good target systems).

C. (U) FOREIGN ASSESSMENT SUPPORT OBJECTIVES

(S/NF) From a research perspective, the objective of foreign assessment is to determine the degree to which claims from foreign laboratories can be confirmed in a U.S.-based setting. In science, replication is critical for understanding.

V. (U) <u>BASIC RESEARCH PLAN FOR ANOMALOUS COGNITION</u>:

A. (U) BASIC APPROACH

(S/NF) The link of basic and applied research with other applications investigations or with research activities is shown on Figure 2. The top of the chart shows that for any research or application task, certain conditions must be met (e.g., a reliable calibrated individual is required; proper scientific procedures need to be developed, etc.). Once these basic foundations are laid, then basic/applied research can be initiated with a reasonable expectation of success and with assurance that results will not be ambiguous or fail scientific scrutiny.

(S/NF) This chart also illustrates the difference between basic and applied research; applied research relates to various methods for collecting, recording, improving and analyzing data output, while basic research is aimed at phenomena understanding. In this chart, the "detector" is the human brain/mind, the "source" is the target or an aspect of the target, and "transmission" refers to notions of how information and/or energy are actually transmitted between source and detector.

(U) Figure 3 illustrates the interdisciplinary scope that will be brought to bear on this research problem. Leading-edge researchers in their various fields can provide clues, if not make direct contributions, that will assist in phenomena and applications understanding. Appendix C lists candidate research support facilities that could be involved in this long-range

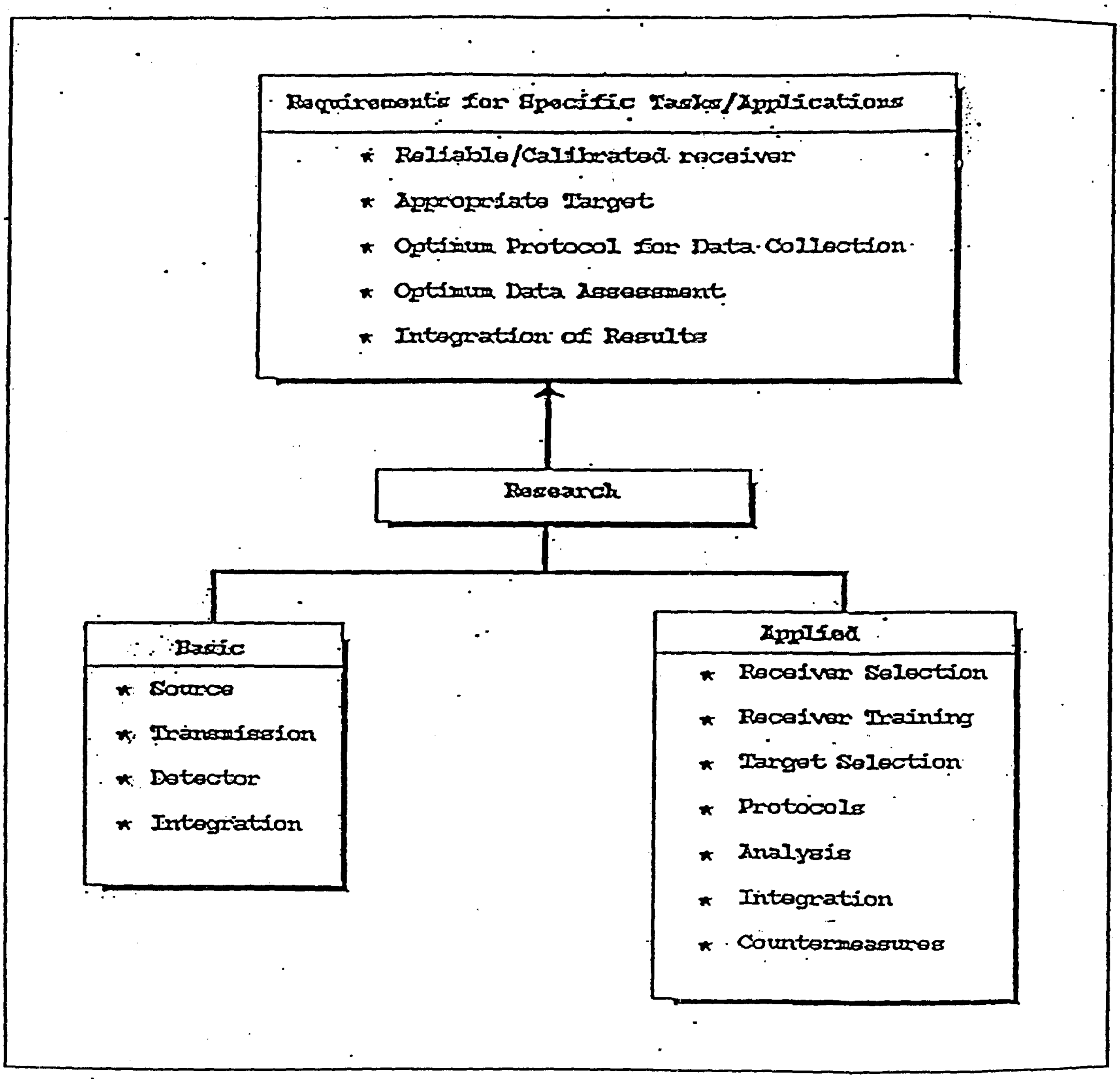

Figure 2 (U) Research Objectives

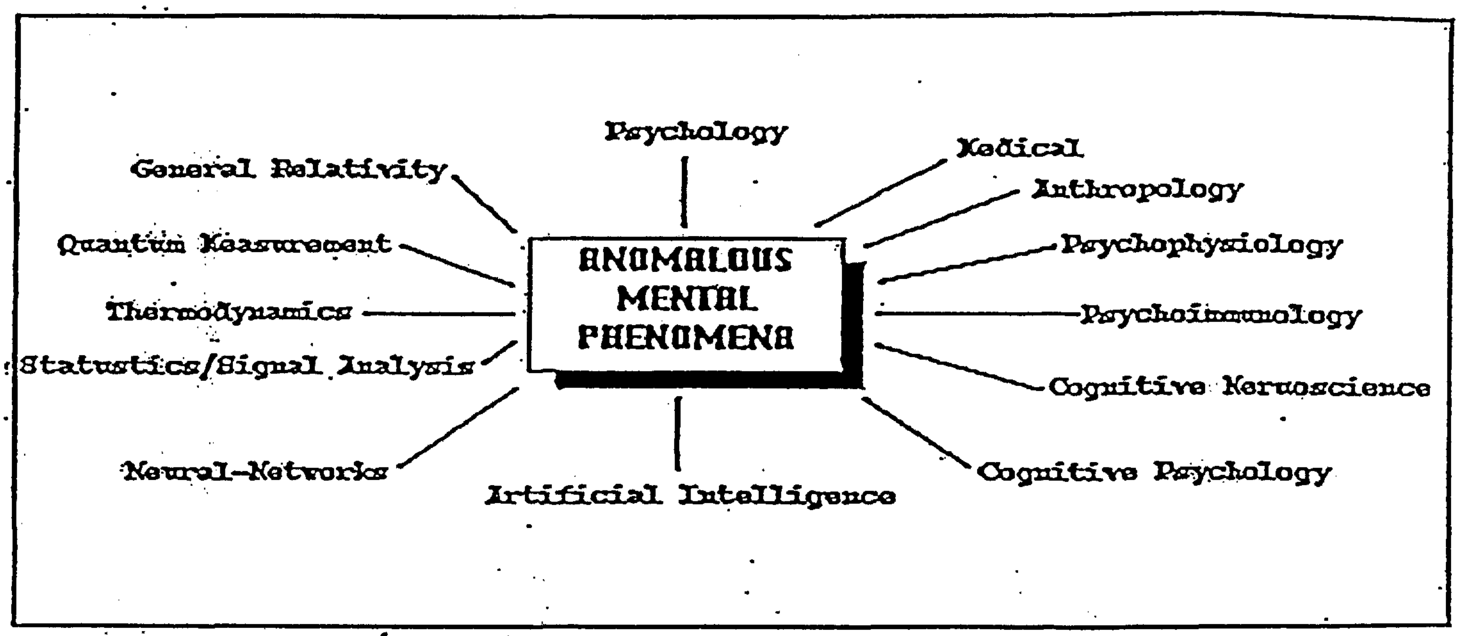

Figure 3 (U) Integration of Scientific Disciplines

effort. Appendix D outlines pertinent research literature applicable to this field. Final selection will be based on how well the activities if these institutions will fit into specific time-lines and priorities to be established in FY95. Figure 4 lists milestones for the anomalous cognition basic research to be conducted under this plan.

B. (U) RESEARCH DETAILS

1. (U) Source.

(S/NF/SG/LIMDIS) Source research will address those topics that show promise for understanding the characteristics of the target or target area that may play a role in anomalous cognition (AC) occurrence and data quality. Aspects of the target that can be defined by conventional information theory (involving entropy/information content) will be explored in-depth. A wide variety of targets with a wide range of information content, dynamics, or other parameters will be examined to explore this possible link. If not successful, other approaches to investigate the targets' innate nature and its possible link to phenomenon occurrence will be initiated. Definitive data in this area would also have implications for defining those targets which have the highest probability of successful data acquisition in an operational setting, thus establishing operational tasking parameters.

2. (U) Transmission.

(S/NF) The pursuit of possible transmission mechanisms for AC phenomena is essentially the most significant basic research task and also the most difficult to formulate. In this effort, a theoretical basis will be developed from extensions of current theory in light of recent advanced physics formulations. Some of these formulations permit unusual "information flows" that may, in fact, have relevance for this phenomenon. Testable models/constructs will be developed and evaluated. A variety of other possible explanations involving extensions of gravitation theory, quantum physics or other areas will be constructed and tested where possible. Some of these tests may require close cooperation of leading-edge researchers using equipment in their facility.

(C/NF) Effort in this area will also focus on integrating diverse aspects of the source, transmission, and detector categories. For example, it will examine how "targeting" occurs. Insight will be drawn from in-depth reviews of various unusual physical effects identified by physical sciences researches. These include distant particle coupling (Bell's theorem), ideas from quantum gravity, possible electrostatic/gravity interactions, unusual quantum physics,

UNCLASSIFIED

ACTIVITY	TIME FRAME: 1995 | 1996 | 1997 | 1998 | 1999
SOURCE RESEARCH (TARGET)	Information/Entropy Analysis Various Target Attributes (Size, Form, Content)
TRANSMISSION RESEARCH (MECHANISM)	Four-Dimensional Calculations (Relativity Extensions) Unconventional Waves (Laboratory) (Long-Range Tests) Variables (Distance, Shielding, Energy)
DETECTOR RESEARCH (BRAIN)	Neuroscience (EEG, Memory, Etc.) Environmental Factors Other Physiology (Electrical, Infrared) Implications from Medical/Animal Research
INTEGRATION	Physical Sciences (Physics, Statistics, Parallel Processing, Etc.) Psychological Sciences (Psychology, Anthropology, Cognitive, Mental, Subliminal Perception, Etc.) Medical (Genetics, Etc.)

FIGURE 4 (U) BASIC RESEARCH MILESTONES - ANOMALOUS COGNITON

UNCLASSIFIED

observational theories, vacuum "energy" potential, and a variety of other concepts.

(S/NF) Perhaps the most promising exploratory model of all is one based on little-understood aspects of the fundamental equations for electromagnetic wave propagation (Maxwell's equations). These equations indicate that forms of "wave propagation" could also exist that do not have the conventional electric or magnetic field components (i.e., vector and scalar waves). These waves would not be blocked by matter and therefore could be leading candidates for AC propagation or for certain aspects of AC phenomenon. Research papers [redacted] SG1B SG1B indicate that these waves are considered a leading candidate for AC transmissions by their researchers. Pilot study investigations in this area were conducted by PAG-TA in FY92 with promising preliminary results. Future research could couple with other DIA exploratory R&D efforts in this area currently being explored.

(S/NF/SG/LIMDIS) Research on this topic will be closely integrated with research involving the anomalous phenomena (AP) aspect, since findings in the AP area would have direct implications for phenomena transmission mechanisms in general. Findings from the target (or target source) research area would also provide insight into possible transmission mechanisms. For example, different forms of the same target (e.g., target size, 2D vs 3D, holographic representations) may show patterns in the AC data that might provide clues regarding phenomena mechanisms.

3. (U) <u>Detector</u>.

(U) The most important and promising aspect of understanding the nature of the AC detection system in humans is through modern advances of the neuroscience. Earlier neurophysiological results obtained from magnetoencephalograph (MEG) measurements begun in FY92 will be validated and expanded. This earlier work indicated MEG correlations between visual evoked responses areas of the brain may exist, and that remote stimuli might also be detectable in MEG data. Some of the specific investigations will examine a variety of near and far-field situations, other sensory modes and different types of individuals in order to search for potential variables. It might be possible, with advanced MEG instrumentation, to actually locate the exact brain areas involved in AC phenomena occurrence. Future research in this area could couple with research currently being explored at the National Laboratory.

(U) Other physical/psychophysical aspects of the central nervous system (CNS) will also be explored to look for possible correlates. This would include galvanic skin responses

(GSR) or other parameters.

(U) Related to this overall area are several investigations that relate to possible environmental interactions with the brain that could affect AC data. This would include possible geomagnetic or electromagnetic influences.

(S/NF) A spin-off from findings in this basic research area could be for unique communication applications. MEG correlates might exist between remotely located people. If so, the possibility of transmission of remote messages (via a type of code) might be possible. Since AC phenomenon is not degraded by distance or shielding, the potential of transmitting basic "messages" to individuals in submarines would exist. Preliminary exploration of this application by PAG-TA has yielded promising results.

(S/NF) Another potential spin-off benefit from detector research in this program is that new insights into brain memory or parallel processing might be achieved. This could lead to new directions in advanced computer developments involving neural networks. For example, recent [redacted] indicates that SG1B "wave-like" brain activity occurs in addition to usual neuronal processes. This wave-like phenomenon may have some link to the "phase shift" observed in MEG data from the previous MEG project. Further MEG work involving remote stimuli may help clarify such issues.

4. (U) Integration.

(U) The basic research activities will liberally avail itself of the existing research communities that specialize in neuroscience, physics and statistics and the broader psychological/social sciences. Direct support with a variety of university departments, national and international, will be explored. PAG-TA contacts with such national laboratories as Los Alamos, Lawrence Livermore, Oak Ridge, and have indicated an interest on their part in supporting the research efforts. Frequent conferences and data exchanges are anticipated. These data exchanges will insure that a proper interdisciplinary approach is maintained, and that findings from other disciplines will be incorporated in this program where appropriate. This peer group dialogue will greatly benefit research sponsored through this plan, new ideas will be generated, and possibly clues regarding phenomena operation will be easier to identify.

(U) Some specific interdisciplinary examples that will benefit this program are as follows:

- In 1990 The American Anthropological Association (AAA) formed a new division, the Society for the

Anthropology of Consciousness (SAC). This division has established a technical journal to support interdisciplinary, cross-cultural, experimental, and theoretical approaches to the study of consciousness. This group may be able to contribute this program by providing cross-cultural examples. These members might also assist in the assessment of foreign data in this area.

- The psychophysiology of vision has already contributed to the earlier program. This plan calls for a collaborative effort with researcher in an attempt to understand how the central nervous system process subliminal stimuli. This should assist in understanding how MEG correlates occur.

- The relationship between mind and body is currently discussed in the research literature as well as in the popular press. Researcher at the California Institute for Transpersonal Psychology (CITP) have been active in investigating the role of mental attitudes and body chemistry. While there may not be a direct link with AC, and exchange of techniques and experimental designs would be helpful.

- The Journal of Cognitive Neuroscience contains at least one article of interest in each issue. This discipline is where most of the cognitive work with the neuromagnetism is conducted. There is the possibility of joint investigations with researchers performing MEG investigations at the National Institutes of Health (NIH).

- Stanford University has been conducting research on internal mental imagery. The manipulation and control of this imagery is extremely important in understanding the source of internal noise during an AC session. A collaborative effort with Stanford should lead to methods for noise reduction.

- Neural networks are particularly good at recognizing subtle patterns in complex data, and are being applied in the subjective arena of decision making in business. In order to improve AC analysis, the program will conduct a collaborative effort with scientists who are active in neural network research and with selected individuals who have had success with interpreting highly subjective data.

- Statistics is the heart of AC research in that most of the results are usually quoted in statistical terms. Hypothesis testing has traditionally been the primary focus, but there are other possible approaches that should be explored. Statistics researchers at Harvard have already expressed interest in contributing to the research effort.

- A major portion of the effort will be a search for a AC evoked response in the brain. Sophisticated processing is required in that magnetic signals from the brain can not be easily characterized by standard statistical practices. Several research facilities can contribute.

- Classical statistical thermodynamics may be the heart of understanding the nature of an AC source of information. A physical property called entropy may be related to what is sensed by AC. The program intends to collaborate with a variety of university physics departments to calculate the appropriate parameters.

(S/NF) The specific experiments to be conducted in these research domains will be defined during the first six to nine months of the program utilizing the recommendations of the working groups mentioned above subject to approval by the Scientific Oversight Committee.

VI. (U) <u>BASIC RESEARCH PLAN FOR ANOMALOUS PERTURBATION</u>:

(S/NF) Figure 5 illustrates the basic approach for investigations "energetics", or anomalous perturbation (AP) phenomenon. Intelligence reporting indicates that this aspect of SG1B AMP [redacted] should receive attention in this research plan to prevent technological surprise. Thus, beginning in FY95, acceptance criteria will be establish with which to judge the historical literature for potential AP effects. Using those criteria, a detailed review of the literature will begin in mid FY95 and considering the size of that data base will continue through FY95. Knowledge gained from this review may provide insights for the development of new AP target systems or provide data so that particular experiments can be replicated. Given the complexity of most AP experiments, considerable time is needed to plan and conduct them properly. If the results warrant, then application development may begin as early as FY96; however the primary task of basic research of AP is to attempt to validate its existence. Findings from foreign research will be examined and factored into this activity as appropriate.

(S/NF) The keys to investigating this area will be in appropriate personnel selection and, very likely, in proper selection of the AP test device. Thus, the initial phase of this effort will involve identification and solicitation of individuals known or claimed to have such talents. For example, certain expert martial arts or yoga practitioners might do well in such experiments due to their strong mental conditioning and ability for intense mental focus. After locating such individuals, various instruments, such as microcomputer devices, sensitive electronic/sensor devices, or other unique or sensitive equipment would be used as targets in AP experiments.

ACTIVITY	TIME FRAME				
	1995	1996	1997	1998	1999
DEVELOP EVALUATION CRITERIA					
PERFORM ANALYSIS	Historical Data Base				
EXAMINE TARGET SYSTEMS		Various Technical Targets Laboratory Setting			
CONDUCT VALIDATION EXPERIMENTS		Advanced Sensors	Complex Components		
PURSUE APPLICATIONS			Far-Field Effects (Countermeasures)		
PERSONNEL SELECTION	Solicit Known Talent		Screening/Training (Develop)		

Figure 5 (U) Basic Research Milestones - Anomalous Perturbation (To Include Biological Systems)

(S/NF) Some of the unique sensor candidates include devices that are highly sensitive to very weak gravitational effects (such as Mossbauer devices or atomic clocks). Perhaps the most promising device is one that involves detection of an unusual non-electromagnetic wave (A vector/scalar wave). If experiments with such sensors are successful, then significant understanding of AP or AC phenomenon would occur. Experiments with such a device is a distinct near-term possibility; consequently this will be given high priority in the early part of this long-range program.

(S/NF) Should these pilot experiments prove successful, then a near and distant experiments would be developed for a wide variety of devices to evaluate application aspects. Potential applications could include, for example, remote switching (in a communication role) or possibly as a countermeasure to minimize effectiveness of threat systems such as sensitive computer components or sensors. Similarly, if these results are successful, they would provide insight regarding potential threats to U.S. systems or security.

(S/NF) The specific experiments to be conducted in these research domains will be defined during the first six to nine months of the program utilizing the recommendations of the working groups mentioned above subject to approval by the Scientific Oversight Committee.

VII. (U) <u>APPLIED RESEARCH PLAN FOR ANOMALOUS COGNITION</u>:

(U) Figure 6 illustrates the overall plan for the applied research portion for several main functional categories.

a. (U) SELECTION

(C) The most promising potential for selecting individuals is to identify ancillary activity that correlates with AC ability. If such a procedure can be identified, then receiver selection can be incorporated as part of other screening tests (e.g., fighter pilot candidacy), and thus large populations can be used. Among the items that will be examined are physiology (e.g., responses of the brain to external stimuli) and hypnotic susceptibility (i.e., an individuals predisposition for being hypnotized). The results of this effort will be examined continuously; however, a decision to end the investigation will occur in mid FY96. Should the results at that time warrant, then refining of the techniques will continue to the end of FY 1998. The reason the initial research spans several years is that to validate even one psychological finding requires long-term testing of candidate individuals. Current statistical methods

ACTIVITY	TIME FRAME: 1995 – 1996 – 1997 – 1998 – 1999
PERSONNEL SELECTION RESEARCH	State Parameters (Hypnosis, Physiology, Etc.) Psychology (Self Report, Behavioral Measures, Etc.) Solicit Known Talent Empirical (Mass Screening)
PERSONNEL TRAINING RESEARCH	State Parameters (Altered States, Subliminal Threshold Measures, Etc.) Empirical Evaluation Practical Application Tests (Increasing Project Difficulty)
APPLICATION EVALUATION RESEARCH	Target Characteristics (Entropy, Size, Etc.) Other Aspects (Target Function, Dynamics, Degree of Importance, Etc.)
PROTOCOL DEVELOPMENT	Operational Conditions (Targets, Feedback, Etc.) Search/Location Projects New Applications/Procedures
ANALYSIS METHOD DEVELOPMENT	Response Definition (Written, Drawn, Physiological Measures, Etc.) Artificial Intelligence (Fuzzy Sets, Etc.) Neural Network Analogies Combination of Methods
DATA INTEGRATION/ ASSIMILATION DEVELOPMENT	Intelligence Data Fusion Methods Training/Seminars Advanced Training Various Customers

Figure 6 (U) Applied Research Milestones - Anomalous Cognition

require many AC sessions, and experience has shown that only a few sessions can be conducted per week for any single individual.

(C) The previous program was able to estimate that approximately one percent of the general population possessed a high-quality, natural AC ability. Because the empirical method (i.e., asking large groups to attempt AC) is labor intensive and very inefficient, it is included in the research plan only as an alternate approach.

b. (U) TRAINING

(S/NF) Training has been a major part of the previous program; however, results of training approaches have been difficult to evaluate and have not been examined systematically. Systematic review of this issue was begun in FY 92. One of the methods that will be examined involves lowering an individual's visual subliminal threshold (i.e., the level below which an individual is not consciously aware of visual material). This could enhance the individual's sensitivity to AC data. Other forms of altered states, such as dreaming and hypnosis, will also be evaluated to see if such states can enhance AC data quality.

(U) Results on these issues should be available at the close of FY95. If no progress has been observed and if there have been no positive results from the basic research, the task ends. However, should any of the variables examined appear promising then the task will be continued.

(S/NF) It is anticipated that all laboratory successes must be validated by simulating operational tasks. These experiments involve identifying the specialty to be tested, the acceptance criteria, and conducting sessions in which the complete target systems are know. This three-year activity runs concurrently with the other tasks but with a one-year offset to allow for planning.

c. (U) TARGET/APPLICATION SELECTION

(C) Based on earlier research, the most promising approach to target selection appears to be a single physical characteristic called entropy (i.e., a measure of inherent target information). Beginning in FY95, two and one half years have been allocated for the detailed study of this aspect of target properties. Initially, little experimentation is required; rather, a retrospective examination of previous target systems should indicate if this approach is valid. Included in this examination are detailed calculations of the information content of natural target scenes.

(S/NF) Beginning in mid FY96, other potential intrinsic target properties will be examined. For example, a target may be more readily sensed by AC if the collection of elements at the site (e.g., landmark, buildings, roads) constitute a conceptually coherent unit as opposed to a collage of unrelated items. Quantitative definition of targets will also be developed that include non-physical target parameters such as function, meaning, or relationships. These aspects are highly important in most operational projects and need to be quantified.

(S/NF) Part of this effort will involve investigations that serve two purposes: (1) add insight into the phenomenon; and (2) help evaluate the feasibility of certain potential applications. For example, long distance experiments could be conducted to or from deep caves or submarines in deep water to test communication potential and transmission theories. Experiments could also be conducted to targets on board space platforms to test distance and gravitational effects. Experiments to or from magnetically shielded rooms or certain earth locations (e.g., the magnetic pole) might indicate if magnetic fields influence the phenomenon. Experiments to opposite sides of the earth might also indicate if a mass or gravity effect can be noted.

(S/NF/SG/LIMDIS) This area of investigation will be integrated with a variety of applications in coordination with findings/investigations pursued by the in-house effort. Figure 9 identifies the main application or operational areas. Along with types of data desired. This activity will be integrated, where possible, into in-house pursuits that will explore these areas in a systematic fashion. Initial emphasis will be in counternarcotics and counterterrorism areas.

(S/NF/SG/LIMDIS) Specific types of applications that will be explored in-depth include the search problem. Search tasks are expected to remain as high priority operational tasks (e.g., hostage location, lost equipment or system location). Search tasks are complicated by timing issues, especially if the missing target is being moved frequently. Related to this will be examination of predictive capability in order to evaluate feasibility of detecting hostile plans and intentions in advance. Pilot studies of other areas (e.g., code breaking, medical diagnostics, low intensity conflict support) will also be initiated.

(S/NF/SG/LIMDIS) Another application area that will be examined is "communications". Previous research indicates that with proper protocols, basic or coded messages can be sent and received via AC procedures. Redundant coding methods can readily enhance probability of success, and new statistical methods can also improve success rates. Communication applications may have significant value for search problems by providing additional information on location of kidnapped or hostage victims. Such techniques might also help in determining hostage or POW state-of-health or other significant issues.

d. (U) PROTOCOLS

(U) Given the laboratory success of AC experimentation, the protocol task can build upon a substantial literature. Determining optimal, specialty-dependent protocols only require extending current concepts. Several years are required due to the statistical nature of analysis that is required to determine the effects of environment, receiver, target and feedback conditions. Several high-interest application areas (such as search/location) will be examined in detail. A variety of session procedures will be evaluated to determine those that are beneficial to improving data quality.

(S/NF) Protocol effectiveness may be measured by quality, quantity, and/or usefulness of the AC information elicited by its use. The requirements for protocols that are designed for laboratory settings are considerably more restrictive than those required for operational settings. For example, providing limited information to a receiver while an operational session is in progress (i.e., intermediate feedback) might facilitate the acquisition of the desired data. This kind of feedback is strictly prohibited, however, in most protocols designed for laboratory experiments. Protocols may also vary depending on nature of the data required. For example, for some search projects, only general data may be adequate. For such cases would not require development of highly specific details and protocols the sessions would not be as complex.

(U) A detailed protocol will need to consider a variety of potential session variables such as the individuals' physical environment, mental state and attitude, and how the target or task is designated (e.g., coordinates, abstract terms). Other data includes specifics of the session (monitor present or not), type of feedback, type of response data (e.g., predictive), and mode and method of response (e.g., drawings, verbal).

(S/NF) Concurrently, the only known way to resolve the above issues is to conduct a large number of trials for a given individual with as many of the potential variables as possible held constant. Standard statistical methods can then be used to identify trends, patterns, and operational constraints.

e. (U) DATA ANALYSIS

(U) This area requires extensive review of leading analysis tools, such as those required for describing imprecise concepts or data (i.e., artificial intelligence techniques, fuzzy sets). This work will be combined with

findings from neural network analysis and research, or possibly combinations of other emerging advanced analysis methods.

(S/NF) Various approaches that are anticipated to directly benefit operational evaluations. One promising technique involves procedures based on an adaptive (frequent data base update) approach. This will permit an individual's progression, and possibly time dependent data variables in an individual's track record, to be identified.

(S/NF) In addition to the search for new analysis methods, the current methods will also be reexamined. Laboratory requirements differ from those for operational activities in that the target can be controlled and well defined. For operational activities, uncertainties in tasking may arise, especially if operational requirements are changing or if some of the initial "known" data are incorrect. Such uncertainties complicate later analyses.

(S/NF) Analysis methods will also be developed that can make predictions on data quality for any given task. This will require development of an extensive track record for each individual based on both controlled and operational projects.

(S/NF) These analysis methods will also address certain practical issues. For example, a detailed, high-quality example of AC data may have little value to an intelligence analyst if that information was known from other sources. Likewise, a poor example of AC data might provide a single element as a tip-off for other assets, or provide the missing piece in a complex analysis, and thus be quite valuable. The intelligence utility of AC data may in some cases be only weakly connected to the AC quality. Therefore a data fusion analysis procedure is needed for AC-derived operational data. Methods that permit appropriate data analysis from an accuracy and utility viewpoint will be developed.

f. (U) INTEGRATION

(U) This activity would be an on-going review/ integration effort in order to identify patterns or clues useful for understanding practical aspects of this phenomenological area.

(S/NF) Identifying approaches and procedures that permit assimilation of AC data from operational support projects into all-source intelligence analysis procedures will also be part of this support activity. Depending on results of applied research findings and operational pursuits, a basic seminar/training program for other applications-oriented elements might be established. Such a training/seminar program would focus on basic techniques and would augment possible operational training activity that might become part of the in-house effort. This would require several years to develop and establish.

(S/NF) The specific experiments to be conducted in these research domains will be defined during the first six to
SG1B nine months of the program utilizing the recommendations of the working groups mentioned above subject to approval by the Scientific Oversight Committee.

SG1B

Approved For Release 2003/04/18 : CIA-RDP96-00789R002700010001-1

IX. (U) <u>POTENTIAL RESEARCH RETURN</u>:

(S/NF/SG/LIMDIS) The research pursuits identified in the overall research and peer review plan have the potential for achieving highly significant results using AMP to address problems of national security by pushing the phenomena to their natural limits. This overall result can be achieved by accomplishing the aforementioned program plan goals.

X. (U) <u>PROGRAM OVERSIGHT</u>

A. (U) PROJECT OVERSIGHT METHODOLOGY:

1. (U) PROGRAM MANAGEMENT/OVERSIGHT

(S/NF) DIA, as executive agent, proposes to implement a management structure that fosters a proactive, responsive, and creative environment for this activity. Both the external research and in-house activities will be centered in the Technology Assessment and Support Activity under the supervision of the Chief, Office for Ground Forces (DIA/PAG).

2. (U) SCIENTIFIC OVERSIGHT

(S/NF) Scientific oversight will be provided by the SEP.

3. (U) CONTRACTOR OVERSIGHT

a. (U) A contractor sponsored Scientific Oversight Committee (SOC), consisting of scientists from the following disciplines: physics, astronomy, statistics, neuroscience, and psychology, will be tasked with the following:

-- (U) Reviewing and approving all

experimental protocols prior to the collection of experimental data.

-- (U) Reviewing all experimental final reports as if they were submissions to technical scientific journals.

-- (U) Proposing directions for further research.

-- (U) Conducting un-announced drop-in privileges to view experiments in progress.

b. (U) An contractor sponsored Human Use Review Board will also be formed and charged with the responsibility of assuring compliance with all U.S. and DoD regulations with regard to the use of humans in experimentation and assuring their safety. Members should represent the health, legal, and spiritual professions IAW government guidelines.

XI. (U) DEVELOPMENT OF EVALUATION CRITERIA:

A. (U) SCIENTIFIC VALIDITY

(S/NF) A thorough review of DoD's activities in AMP was conducted in 1987 to evaluate the use of AMP for intelligence gathering purposes. The overall findings of this evaluation were that "...the Project Review Group has determined to its satisfaction that the work of the Enhanced Human Performance Group is scientifically sound...and is providing valuable insight into the nature of an anomaly which have a significant impact on the DoD." This research and development program will both draw from and add to this extensive data base to further demonstrate the scientific validity and practicality of AMP.

B. (U) PERFORMANCE

(S/NF) The ability of the STAR GATE program to produce results that have an intelligence value can only be measured by customer feedback evaluations. STAR GATE has developed feedback mechanisms and procedures for customers that should result in a method of quantifying this subjective feedback data so that operational value added and cost-effectiveness can be measured.

XII. (U) BUDGET AND RESOURCE REQUIREMENTS (FYs 95-99):

(S/NF/SG/LIMDIS) Due to the diversity of the STAR GATE mission/objectives, both external resources and in-house expertise are required. Since this Activity possesses no in-house R&D capability, an absolute need for external R&D support is required to meet Congressional concerns which are addressed in this program plan. A balance will be maintained between external and in-house activities, and every effort will be made to integrate and link these activities where appropriate. The external aspect permits a wide range of expertise covering many disciplines to be focused on this area; this also has the benefit of ensuring peer group review and of facilitating a variety of scientific interactions. In-house personnel with a wide-range of expertise in this phenemenology will need to be retained to make this proposed plan work.

(S/NF/SG/LIMDIS) In order to fulfill Congressional Direction, the DIA proposes to convene a Scientific Evaluation Panel (SEP) composed of representatives from each of the Service Scientific Advisory Boards. The purpose of the SEP is to review and validate the methodology outlined in the plan in order to address the cost-effectiveness and performance criteria for the STAR GATE program's research and development objectives and to propose recommendations as to which objectives should be pursued and the program scope required to achieve those objectives. If the SEP determines that objectives in the plan are viable and executable, the General Defense Intelligence Program (GDIP) Manager will complete this initiative with others for limited available resources remaining in the program.

(U) The proposed ongoing R&D effort will be reviewed every two years by the SEP to determine whether the STAR GATE program can show results that are cost-effective and satisfy reasonable performance criteria.

(C) An annual report will document the current operational, technical and administrative status of the program.

APPENDIX A

CONGRESSIONALLY-DIRECTED ACTION

DEFENSE AUTHORIZATION CONFERENCE

(S/NF) REQUEST: "The conferees are concerned that insufficient funds have been spent on research and development to establish the scientific basis for the STAR GATE program. The conferees direct the Director of DIA to prepare a program plan and to submit an appropriate budget request for a research effort, over several years, to determine whether the STAR GATE program can show results that are cost-effective and satisfy reasonable performance criteria. This plan, and any research under this program, should be subject to peer review by neutral scientific experts. The Director of DIA is directed to prepare this research and peer review plan within existing program funds."

APPENDIX B

TERMINOLOGY AND DEFINITIONS

(U) PHENOMENA TERMINOLOGY:

(U) This phenomenological area has had a variety of descriptive terms over the years, such as paranormal, parapsychological, or as psychical research. Foreign researchers use other terms: "psychoenergetics" in the USSR; "extraordinary human function" in the People's Republic of China (PRC). In general, this field is concerned with a largely unexplored area of human consciousness/subconsciousness interactions associated with unusual or underdeveloped human capabilities.

(U) Recently, researchers have shown a preference for terms that are neutral and that emphasizes the anomalous or enigmatic nature of this phenomena. The term anomalous mental phenomena (AMP), is generally preferred.

(U) This area has two aspects; information access and energetics influence. Information access refers to a mental ability to describe remote areas or to access concealed data that are otherwise shielded from all known sensory channels. A recent term for this ability is anomalous cognition (AC). This term places emphasis on potential understanding that might be available from advances in sensory/brain functioning research or other related research. Older terms for this aspect have included extra-sensory perception (ESP), remote viewing (RV), and in some cases, precognition.

(U) The energetics aspect refers to the ability to influence, via mental volition, physical or biological systems by an as yet unknown physical mechanism. An example of physical system influence would include affecting the output of sensors or electronic devices; biological systems influence would include affecting physiological parameters of an individual. A recent descriptive term for this ability is anomalous perturbation (AP). Older terms for this phenomenon included psychokinesis (PK) or telekinesis.

(U) GENERAL DEFINITIONS:

(S/NF) For this program, basic research is defined to mean any investigation or experiment for determining fundamental processes or for uncovering underlying parameters that are involved in this phenomenon. Basic research is primarily oriented toward understanding the physical, physiological , and psychological mechanisms of anomalous mental phenomena (AMP).

(S/NF) Applied research refers to any investigation directed toward developing particular applications or for improving data quality and reliability. For anomalous cognition (AC) phenomenon, research is primarily directed toward improving the output quality of AC data. This would include ways to develop/improve utility of AC data for variety of potential application. For example, examination of spatial and temporal relationships of AC data could assist in developing a reliable search capability useful for locating missing people or equipment.

APPENDIX C

POTENTIAL RESEARCH SUPPORT FACILITIES

ANOMALOUS MENTAL PHENOMENA

Science Applications International Corp.	Los Altos, CA
Mind Science Foundation	San Antonio, TX
Princeton Engineering Anomalies Laboratory	Princeton Univ, NJ
American Society for Psychical Research	New York, NY
St. John's University	Long Island, NY
Foundation for Research into the Nature of Man	Durham, NC
ARE/Atlantic University	Virginia Beach, VA
University of Virginia	Charlottesville, VA
Psychophysical Research Laboratories	Edinburgh, Scotland
Edinburgh University	Edinburgh, Scotland

OTHER RELATED DISCIPLINES

Psychology	
Stanford University	Stanford, CA
Cornell University	Ithaca, NY
Anthropology	
University of California	Berkeley, CA
University of Arizona	Tucson, AZ
Psychophysiology	
SRI International	Menlo Park, CA
Langly-Portor Neuropsychiatric Institute	San Francisco, CA
Menninger Foundation	Topeka, KS
Psychoimmunology	
California Institute for Transpersonal Psychology	Menlo Park, CA
Cognitive Neuroscience	
Los Alamos National Laboratory	Los Alamos, NM
Sandia National Laboratory	Albuquerque, NM
University of California	San Diego, CA

Cognitive Psychology

Psychology Department, Princeton Univ	Princeton, NJ
Psychology Department, City College of New York	New York, NY

Artificial Intelligence

Massachusetts Institute of Technology	Cambridge, MA
Stanford University	Stanford, CA

Neural Networks

Massachusetts Institute of Technology	Cambridge, MA
Science Applications International Corp	Los Altos, CA

Statistics/Signal Analysis

University of California	Davis, CA
Harvard University	Cambridge, MA

Thermodynamics

Rochester University	Rochester, NY
Physics Department, Stanford University	Stanford, CA

Quantum Measurement

International Business Machines, Research Laboratories	College Park, MD

General Relativity

California Institute of Technology	Pasadena, CA
University of Texas at Austin	Austin, TX

Electromagnetic/Basic Research

Electronetics Corp	Buffalo, NY
Battelle Corp	Columbus, OH
Institute for Advanced Study	Austin, TX

APPENDIX D

RESOURCE LITERATURE

1. A.R.E. Journal
2. Abnormal hypnotic Phenomena
3. American Anthropologist
4. American Ethnologist
5. American Journal of Clinical Hypnosis
6. American Journal of Physiology
7. American Journal of Sociology
8. American Psychologist
9. American Society for Psychical Research
10. Annals of Eugenics
11. Annals of Mathematical Statistics
12. Annales de Sciences Psychiques
13. Archivo di Psicologica Neurologia e Psychiatra
14. Association for the Anthropological Study of Consciousness Newsletter
15. Behavioral and Brain Science
16. Behavioral Science
17. Bell System Technical Journal
18. Biological Psychiatry
19. Biological Review
20. British Journal for the Philosophy of Science
21. British Journal of Psychology
22. Bulletin of the American Physical Research
23. Bulletin of the Boston Society for Psychic Research
24. Bulletin of the Los Angeles Neurological Societies
25. Contributions to Asian Studies
26. Electroencephalography and Clinical Neurophysiology
27. Endeavour
28. Ethnology
29. Exceptional Human Experience
30. Experientia
31. Experimental Medicine and Surgery
32. Fate
33. Fields within Fields
34. Foundations of Physics
35. Hibbert Journal
36. Human Biology
37. International Journal of Clinical and Experimental Hypnosis
38. International Journal of Comparative Sociology

39. International Journal of Neuropsychiatry
40. International Journal of Parapsychology
41. International Journal of Psychoanalysis
42. Journal of Abnormal and Social Psychology
43. Journal of Altered States of Consciousness
44. Journal of Applied Physics
45. Journal of Applied Psychology
46. Journal of Asian and African Studies
47. Journal of Biophysical and Biochemical Cytology
48. Journal of Cell Biology
49. Journal of Communication
50. Journal of Comparative and Physiological Psychology
51. Journal of Consulting Psychology
52. Journal of Existential Psychiatry
53. Journal of Experimental Biology
54. Journal of Experimental Psychology
55. Journal of General Psychology
56. Journal of Genetic Psychology
57. Journal of Mind and Behavior
58. Journal of Nervous and Mental Diseases
59. Journal of Personality
60. Journal of Personality and Social Psychology
61. Journal of Research in PSI Phenomena
62. Journal of Scientific Exploration
63. Journal of the American Academy of Psychoanalysis
64. Journal of the London Mathematical Society
65. Journal of the Royal Anthropological Institute of Great Britain and Ireland
66. Metapsichica
67. Mind-Brain Bulletin
68. Motivation and Emotion
69. Nature
70. Naturwissenschaftliche Rundschau
71. New Horizons
72. New Scientist
73. New Sense bulletin
74. Newsletter of the Parapsychology Foundation
75. Parapsychology Bulletin
76. Parapsychology Abstracts International
77. Parapsychology Review
78. Perceptual and Motor Skills
79. Philosophy of Science
80. Physiology and Behavior
81. Proceedings of the Society for Psychical Research
82. Psychedelic Review
83. Psychic

84. Psychic Science
85. Psychoanalytic Quarterly
86. Psychoanalytic Review
87. Psychological Bulletin
88. Psychometrika
89. Psychophysiology
90. Physics Today
91. Renti Teyigongneng (EFHB Research) [PRC]
92. Revue Metapsychique
93. Revue Philosophique
94. Revue Philosophique de la France et de L'Etranger
95. Revue Philosophique Applique
96. Science
97. Skeptical Inquirer
98. Social Studies of Science
99. Subtle Energies
100. The Humanistic Psychology Institute
101. The Journal of Parapsychology
102. The Journal of the American Society for Psychical Research
103. Theta
104. Tijdschrif voor Parapsychologie
105. Tomorrow
106. Voprosy Filosofi (Questions of Philosophy) [RUSSIA]
107. Western Canadian Journal of Anthropology
108. Zeitschrift fur die Gesamte Neurologie und Psychiatrie
109. Zietschrift fur Parapsychologie und Grenzgebeite der Psychologie
110. Zietschrift fur Tierpsychologie
111. Zietschrift fur Vergleichende Physiologie
112. Zetetic Scholar
113. Zhongguo Shebui Kexue (China Social Sciences) [PRC]
114. Ziran Zazhi (Nature) [PRC]

APPENDIX E
CURRENT CONTRACTOR SCIENTIFIC OVERSIGHT COMMITTEE MEMBERSHIP

Steven A. Hillyard
- Professor of Neurosciences, Department of Neurosciences, University of California, San Diego.
- Author or coauthor of 118 technical neuroscience publications.
- Eighty-two invited presentations at technical conferences.
- Ph.D., Yale University, 1968 (Psychology).

S. James Press
- Professor of Statistics, Department of Statistics, University of California, Riverside.
- Author or coauthor of 132 statistics publications.
- Author of 12 books and/or monographs.
- Ph.D., Stanford University, 1964 (Statistics).

Garrison Rapmund
- Responsible for facilitating transfer of Strategic Defense Initiative technologies to health care industries.
- Major General, USA retired in 1986 as Assistant Surgeon General (R&D) and Commander, Army Medical R & D Command.
- M.D., Columbia University, 1953 (Pediatrics).

Melvin Schwartz
- Associate Director for High Energy and Nuclear Physics, Brookhaven National Laboratory.
- Author or coauthor of 40 technical publications in high energy physics, author of "Principles of Electrodynamics."
- Nobel Prize, Physics (1988).
- Ph.D., Columbia University, 1958 (Physics).

Yervant Terzian
- Professor of Physical Sciences, Chairman of the Department of Astronomy, Cornell University.
- Author/coauthor of numerous technical publications and books.
- Ph.D., Indiana University, 1965 (Astronomy).

Phillip G. Zimbardo
- Professor of Psychology, Department of Psychology, Stanford University.
- Author/coauthor of numerous experimental psychology publications.
- Ph.D., Yale University, 1959 (Psychology).

APPENDIX F
CURRENT CONTRACTOR INSTITUTIONAL REVIEW BOARD MEMBERSHIP

Byron Wm. Brown, Jr., Ph.D.
- Biostatistics, Stanford University

Gary R. Fujimoto, M. D.
- Occupational Medicine, Palo Alto Medical Foundation

John Hanley, M. D.
- Neuropsychiatry, University of California, Los Angeles

Robert B. Livingston, M. D.
- Neuroscience, University of California, San Diego

Robin P. Michelson, M. D.
- Otolaryngology, University of California, San Francisco

Ronald Y. Nakasone, Ph.D.
- Buddhist Studies, Institute of Buddhist Studies, Berkeley, CA

Garrison Rapmund, M. D. (Chair)
- Air Force Science Advisory Board

Louis J. West, M. D.
- Neuropsychiatry, University of California, Los Angeles

APPENDIX G
ACADEMIC STUDIES REGARDING THE SCIENTIFIC VALIDITY OF AMP

Psychological Bulletin (January, 1994)

Version 4.7
October 1, 1993

Does Psi Exist? Replicable Evidence for an Anomalous Process of Information Transfer

Daryl J. Bem and Charles Honorton

Most academic psychologists do not yet accept the existence of psi, anomalous processes of information or energy transfer (such as telepathy or other forms of extrasensory perception) that are currently unexplained in terms of known physical or biological mechanisms. We believe that the replication rates and effect sizes achieved by one particular experimental method, the *ganzfeld* procedure, are now sufficient to warrant bringing this body of data to the attention of the wider psychological community. Competing meta-analyses of the ganzfeld database are reviewed, 1 by R. Hyman (1985), a skeptical critic of psi research, and the other by C. Honorton (1985), a parapsychologist and major contributor to the ganzfeld database. Next the results of 11 new ganzfeld studies that comply with guidelines jointly authored by R. Hyman and C. Honorton (1986) are summarized. Finally, issues of replication and theoretical explanation are discussed.

The term *psi* denotes anomalous processes of information or energy transfer, processes such as telepathy or other forms of extrasensory perception that are currently unexplained in terms of known physical or biological mechanisms. The term is purely descriptive: It neither implies that such anomalous phenomena are paranormal nor connotes anything about their underlying mechanisms.

Does psi exist? Most academic psychologists don't think so. A survey of more than 1,100 college professors in the United States found that 55% of natural scientists, 66% of social scientists (excluding psychologists), and 77% of academics in the arts, humanities, and education believed that ESP is either an established fact or a likely possibility. The comparable figure for psychologists was only 34%. Moreover, an equal number of psychologists declared ESP to be an impossibility, a view expressed by only 2% of all other respondents (Wagner & Monnet, 1979).

Daryl J. Bem, Department of Psychology, Cornell University; Charles Honorton, Department of Psychology, University of Edinburgh, Edninburgh, Scotland.

Sadly, Charles Honorton died of a heart attack on November 4, 1992, 9 days before this article was accepted for publication. He was 46. Parapsychology has lost one of its most valued contributors. I have lost a valued friend.

This collaboration had its origins in a 1983 visit I made to Honorton's Psychophysical Research Laboratories (PRL) in Princeton, New Jersey, as one of several outside consultants brought in to examine the design and implementation of the experimental protocols.

Preparation of this article was supported, in part, by grants to Charles Honorton from the American Society for Psychical Research and the Parapsychology Foundation, both of New York City. The work at PRL summarized in the second half of this article was supported by the James S. McDonnell Foundation of St. Louis, Missouri, and by the John E. Fetzer Foundation of Kalamazoo, Michigan.

Helpful comments on drafts of this article were received from Deborah Delanoy, Edwin May, Donald McCarthy, Robert Morris, John Palmer, Robert Rosenthal, Lee Ross, Jessica Utts, Philip Zimbardo, and two anonymous reviewers.

Correspondence concerning this article should be addressed to Daryl J. Bem, Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853. (Electronic mail may be sent to d.bem@cornell.edu).

Psychologists are probably more skeptical about psi for several reasons. First, we believe that extraordinary claims require extraordinary proof. And although our colleagues from other disciplines would probably agree with this dictum, we are more likely to be familiar with the methodological and statistical requirements for sustaining such claims, as well as with previous claims that failed either to meet those requirements or to survive the test of successful replication. Even for ordinary claims, our conventional statistical criteria are conservative. The sacred $p = .05$ threshold is a constant reminder that it is far more sinful to assert that an effect exists when it does not (the Type I error) than to assert that an effect does not exist when it does (the Type II error).

Second, most of us distinguish sharply between phenomena whose explanations are merely obscure or controversial (e.g., hypnosis) and phenomena such as psi that would appear to fall outside our current explanatory framework altogether. (Some would characterize this as the difference between the unexplained and the inexplicable.) In contrast, many laypersons treat all exotic psychological phenomena as epistemologically equivalent; many even consider déjà vu to be a psychic phenomenon. The blurring of this critical distinction is aided and abetted by the mass media, "new age" books and mind-power courses, and "psychic" entertainers who present both genuine hypnosis and fake "mind reading" in the course of a single performance. Accordingly, most laypersons would not have to revise their conceptual model of reality as radically as we would to assimilate the existence of psi. For us, psi is simply more extraordinary.

Finally, research in cognitive and social psychology has sensitized us to the errors and biases that plague intuitive attempts to draw valid inferences from the data of everyday experience (Gilovich, 1991; Nisbett & Ross, 1980; Tversky & Kahneman, 1971). This leads us to give virtually no probative weight to anecdotal or journalistic reports of psi, the main source cited by our academic colleagues as evidence for their beliefs about psi (Wagner & Monnet, 1979).

Ironically, however, psychologists are probably *not* more familiar than others with recent experimental research on psi. Like most psychological research, parapsychological research is reported primarily in specialized journals; unlike most psychological research, however, contemporary parapsychological research is not usually reviewed or summarized in psychology's textbooks, handbooks, or mainstream journals. For example, only 1 of 64 introductory psychology textbooks recently surveyed even mentions the experimental procedure reviewed in this article, a procedure that has been in widespread use since the early 1970s (Roig, Icochea, & Cuzzucoli, 1991). Other secondary sources for nonspecialists are frequently inaccurate in their descriptions of parapsychological research. (For discussions of this problem, see Child, 1985; and Palmer, Honorton, & Utts, 1989.)

This situation may be changing. Discussions of modern psi research have recently appeared in a widely used introductory textbook (Atkinson, Atkinson, Smith, & Bem, 1990, 1993), two mainstream psychology journals (Child, 1985; Rao & Palmer, 1987), and a scholarly but accessible book for nonspecialists (Broughton, 1991). The purpose of the present article is to supplement these broader treatments with a more detailed, meta-analytic presentation of evidence issuing from a single experimental method: the *ganzfeld* procedure. We believe that the replication rates and effect sizes achieved with this procedure are now sufficient to warrant bringing this body of data to the attention of the wider psychological community.

The Ganzfeld Procedure

By the 1960s, a number of parapsychologists had become dissatisfied with the familiar ESP testing methods pioneered by J. B. Rhine at Duke University in the 1930s. In particular, they believed that the repetitive forced-choice procedure in which a subject repeatedly attempts to select the correct "target" symbol from a set of fixed alternatives failed to capture the circumstances that characterize reported instances of psi in everyday life.

Historically, psi has often been associated with meditation, hypnosis, dreaming, and other naturally occurring or deliberately induced altered states of consciousness. For example, the view that psi phenomena can occur during meditation is expressed in most classical texts on meditative techniques; the belief that hypnosis is a psi-conducive state dates all the way back to the days of early mesmerism (Dingwall, 1968); and cross-cultural surveys indicate that most reported "real-life" psi experiences are mediated through dreams (Green, 1960; Prasad & Stevenson, 1968; L. E. Rhine, 1962; Sannwald, 1959).

There are now reports of experimental evidence consistent with these anecdotal observations. For example, several laboratory investigators have reported that meditation facilitates psi performance (Honorton, 1977). A meta-analysis of 25 experiments on hypnosis and psi conducted between 1945 and 1981 in 10 different laboratories suggests that hypnotic induction may also facilitate psi performance (Schechter, 1984). And dream-mediated psi was reported in a series of experiments conducted at Maimonides Medical Center in New York and published between 1966 and 1972 (Child, 1985; Ullman, Krippner, & Vaughan, 1973).

In the Maimonides dream studies, two subjects—a "receiver" and a "sender"—spent the night in a sleep laboratory. The receiver's brain waves and eye movements were monitored as he or she slept in an isolated room. When the receiver entered a period of REM sleep, the experimenter pressed a buzzer that signaled the sender—under the supervision of a second experimenter—to begin a sending period. The sender would then concentrate on a randomly chosen picture (the "target") with the goal of influencing the content of the receiver's dream.

Toward the end of the REM period, the receiver was awakened and asked to describe any dream just experienced. This procedure was repeated throughout the night with the same target. A transcription of the receiver's dream reports was given to outside judges who blindly rated the similarity of the night's dreams to several pictures, including the target. In some studies, similarity ratings were also obtained from the receivers themselves. Across several variations of the procedure, dreams were judged to be significantly more similar to the target pictures than to the control pictures in the judging sets (failures to replicate the Maimonides results were also reviewed by Child, 1985).

These several lines of evidence suggested a working model of psi in which psi-mediated information is conceptualized as a weak signal that is normally masked by internal somatic and external sensory "noise." By reducing ordinary sensory input, these diverse psi-conducive states are presumed to raise the signal-to-noise ratio, thereby enhancing a person's ability to detect the psi-mediated information (Honorton, 1969, 1977). To test the hypothesis that a reduction of sensory input itself facilitates psi performance, investigators turned to the ganzfeld procedure (Braud, Wood, & Braud, 1975; Honorton & Harper, 1974; Parker, 1975), a procedure originally introduced into experimental psychology during the 1930s to test propositions derived from Gestalt theory (Avant, 1965; Metzger, 1930).

Like the dream studies, the psi ganzfeld procedure has most often been used to test for telepathic communication between a sender and a receiver. The receiver is placed in a reclining chair in an acoustically isolated room. Translucent ping-pong ball halves are taped over the eyes and headphones are placed over the ears; a red floodlight directed toward the eyes produces an undifferentiated visual field and white noise played through the headphones produces an analogous auditory field. It is this homogeneous perceptual environment that is called the *Ganzfeld* ("total field"). To reduce internal somatic "noise," the receiver typically also undergoes a series of progressive relaxation exercises at the beginning of the ganzfeld period.

The sender is sequestered in a separate acoustically isolated room, and a visual stimulus (art print, photograph, or brief videotaped sequence) is randomly selected from a large pool of such stimuli to serve as the target for the session. While the sender concentrates on the target, the receiver provides a continuous verbal report of his or her ongoing imagery and mentation, usually for about 30 minutes. At the completion of the ganzfeld period, the receiver is presented with several stimuli (usually four) and, without knowing which stimulus was the target, is asked to rate the degree to which each matches the imagery and mentation experienced during the ganzfeld period. If the receiver assigns the highest rating to the target stimulus, it is scored as a "hit." Thus, if the experiment uses judging sets containing four stimuli (the target and three decoys or control stimuli), the hit rate expected by chance is .25. The ratings can also be analyzed in other ways; for example, they can be converted to ranks or standardized scores within each set and analyzed parametrically across sessions. And, as with the dream studies, the similarity ratings can also be made by outside judges using transcripts of the receiver's mentation report.

Meta-Analyses of the Ganzfeld Database

In 1985 and 1986, the *Journal of Parapsychology* devoted two entire issues to a critical examination of the ganzfeld database. The 1985 issue comprised two contributions: (a) a meta-analysis and critique by Ray Hyman (1985), a cognitive psychologist and skeptical critic of parapsychological research, and (b) a competing meta-analysis and rejoinder by Charles Honorton (1985), a parapsychologist and major contributor to the ganzfeld database. The 1986 issue contained four commentaries on the Hyman-Honorton exchange, a joint communiqué by Hyman and Honorton, and six additional commentaries on the joint communiqué itself. We summarize the major issues and conclusions here.

Replication Rates

Rates by study. Hyman's meta-analysis covered 42 psi ganzfeld studies reported in 34 separate reports written or published from 1974 through 1981. One of the first problems he discovered in the database was multiple analysis. As noted earlier, it is possible to calculate several indexes of psi performance in a ganzfeld experiment and, furthermore, to subject those indexes to several kinds of statistical treatment. Many investigators reported multiple indexes or applied multiple statistical tests without adjusting the criterion significance level for the number of tests conducted. Worse, some may have "shopped" among the alternatives until finding one that yielded a significantly successful outcome. Honorton agreed that this was a problem.

Accordingly, Honorton applied a uniform test on a common index across all studies from which the pertinent datum could be extracted, regardless of how the investigators had analyzed the data in the original reports. He selected the proportion of hits as the common index because it could be calculated for the largest subset of studies: 28 of the 42 studies. The hit rate is also a conservative index because it discards most of the rating information; a second place ranking—a "near miss"—receives no more credit than a last place ranking. Honorton then calculated the exact binomial probability and its associated z score for each study.

Of the 28 studies, 23 (82%) had positive z scores ($p = 4.6 \times 10^{-4}$, exact binomial test with $p = q = .5$). Twelve of the studies (43%) had z scores that were independently significant at the 5% level ($p = 3.5 \times 10^{-9}$, binomial test with 28 studies, $p = .05$, and $q = .95$), and 7 of the studies (25%) were independently significant at the 1% level ($p = 9.8 \times 10^{-9}$). The composite Stouffer z score across the 28 studies was 6.60 ($p = 2.1 \times 10^{-11}$).[1] A more conservative estimate of significance can be obtained by including 10 additional studies that also used the relevant judging procedure but did not report hit rates. If these studies are assigned a mean z score of zero, the Stouffer z across all 38 studies becomes 5.67 ($p = 7.3 \times 10^{-9}$).

Thus, whether one considers only the studies for which the relevant information is available or includes a null estimate for the additional studies for which the information is not available, the aggregate results cannot reasonably be attributed to chance. And, by design, the cumulative outcome reported here cannot be attributed to the inflation of significance levels through multiple analysis.

Rates by laboratory. One objection to estimates such as those just described is that studies from a common laboratory are not independent of one another (Parker, 1978). Thus, it is possible for one or two investigators to be disproportionately responsible for a high replication rate whereas other, independent investigators are unable to obtain the effect.

The ganzfeld database is vulnerable to this possibility. The 28 studies providing hit rate information were conducted by investigators in 10 different laboratories. One laboratory contributed 9 of the studies, Honorton's own laboratory contributed 5, 2 other laboratories contributed 3 each, 2 contributed 2 each, and the remaining 4 laboratories each contributed 1. Thus, half of the studies were conducted by only 2 laboratories, 1 of them Honorton's own.

Accordingly, Honorton calculated a separate Stouffer z score for each laboratory. Significantly positive outcomes were reported by 6 of the 10 laboratories, and the combined z score across laboratories was 6.16 ($p = 3.6 \times 10^{-10}$). Even if all of the studies conducted by the 2 most prolific laboratories are discarded from the analysis, the Stouffer z across the 8 other laboratories remains significant ($z = 3.67$, $p = 1.2 \times 10^{-4}$). Four of these studies are significant at the 1% level ($p = 9.2 \times 10^{-6}$, binomial test with 14 studies, $p = .01$, and $q = .99$), and each was contributed by a different laboratory. Thus, even though the total number of laboratories in this database is small, most of them have reported significant studies, and the significance of the overall effect does not depend on just one or two of them.

Selective Reporting

In recent years, behavioral scientists have become increasingly aware of the "file-drawer" problem: the likelihood that successful studies are more likely to be published than unsuccessful studies, which are more likely to be consigned to the file drawers of their disappointed investigators (Bozarth & Roberts, 1972; Sterling, 1959). Parapsychologists were among the first to become sensitive to the problem, and, in 1975, the Parapsychological Association Council adopted a policy opposing the selective reporting of positive outcomes. As a consequence, negative findings have been routinely reported at the association's meetings and in its affiliated publications for almost two decades. As has already been shown, more than half of the ganzfeld studies included in the meta-analysis yielded outcomes whose significance falls short of the conventional .05 level.

A variant of the selective reporting problem arises from what Hyman (1985) has termed the "retrospective study." An investigator conducts a small set of exploratory trials. If they yield null results, they remain exploratory and never become part of the official record; if they yield positive results, they are defined as a study after the fact and are submitted for publication. In support of this possibility, Hyman noted that there are more significant studies in the database with fewer than 20 trials than one would expect under the assumption that, all other things being equal, statistical power should increase with the square root of the sample size. Although Honorton questioned the

[1]Stouffer's z is computed by dividing the sum of the z scores for the individual studies by the square root of the number of studies (Rosenthal, 1978).

assumption that "all other things" are in fact equal across the studies and disagreed with Hyman's particular statistical analysis, he agreed that there is an apparent clustering of significant studies with fewer than 20 trials. (Of the complete ganzfeld database of 42 studies, 8 involved fewer than 20 trials, and 6 of those studies reported statistically significant results.)

Because it is impossible, by definition, to know how many unknown studies—exploratory or otherwise—are languishing in file drawers, the major tool for estimating the seriousness of selective reporting problems has become some variant of Rosenthal's file drawer statistic, an estimate of how many unreported studies with z scores of zero would be required to exactly cancel out the significance of the known database (Rosenthal, 1979). For the 28 direct-hit ganzfeld studies alone, this estimate is 423 fugitive studies, a ratio of unreported-to-reported studies of approximately 15:1. When it is recalled that a single ganzfeld session takes over an hour to conduct, it is not surprising that—despite his concern with the retrospective study problem—Hyman concurred with Honorton and other participants in the published debate that selective reporting problems cannot plausibly account for the overall statistical significance of the psi ganzfeld database (Hyman & Honorton, 1986).[2]

Methodological Flaws

If the most frequent criticism of parapsychology is that it has not produced a replicable psi effect, the second most frequent criticism is that many, if not most, psi experiments have inadequate controls and procedural safeguards. A frequent charge is that positive results emerge primarily from initial, poorly controlled studies and then vanish as better controls and safeguards are introduced.

Fortunately, meta-analysis provides a vehicle for empirically evaluating the extent to which methodological flaws may have contributed to artifactual positive outcomes across a set of studies. First, ratings are assigned to each study that index the degree to which particular methodological flaws are or are not present; these ratings are then correlated with the studies' outcomes. Large positive correlations constitute evidence that the observed effect may be artifactual.

In psi research, the most fatal flaws are those that might permit a subject to obtain the target information in normal sensory fashion, either inadvertently or through deliberate cheating. This is called the problem of *sensory leakage*. Another potentially serious flaw is inadequate randomization of target selection.

Sensory leakage. Because the ganzfeld is itself a perceptual isolation procedure, it goes a long way toward eliminating potential sensory leakage during the ganzfeld portion of the session. There are, however, potential channels of sensory leakage after the ganzfeld period. For example, if the experimenter who interacts with the receiver knows the identity of the target, he or she could bias the receiver's similarity ratings in favor of correct identification. Only one study in the database contained this flaw, a study in which subjects actually performed slightly below chance expectation. Second, if the stimulus set given to the receiver for judging contains the actual physical target handled by the sender during the sending period, there might be cues (e.g., fingerprints, smudges, or temperature differences) that could differentiate the target from the decoys. Moreover, the process of transferring the stimulus materials to the receiver's room itself opens up other potential channels of sensory leakage. Although contemporary ganzfeld studies have eliminated both of these possibilities by using duplicate stimulus sets, some of the earlier studies did not.

Independent analyses by Hyman and Honorton agreed that there was no correlation between inadequacies of security against sensory leakage and study outcome. Honorton further reported that if studies that failed to use duplicate stimulus sets were discarded from the analysis, the remaining studies are still highly significant (Stouffer $z = 4.35, p = 6.8 \times 10^{-6}$)

Randomization. In many psi experiments, the issue of target randomization is critical because systematic patterns in inadequately randomized target sequences might be detected by subjects during a session or might match subjects' preexisting response biases. In a ganzfeld study, however, randomization is a much less critical issue because only one target is selected during the session and most subjects serve in only one session. The primary concern is simply that all the stimuli within each judging set be sampled uniformly over the course of the study. Similar considerations govern the second randomization, which takes place after the ganzfeld period and determines the sequence in which the target and decoys are presented to the receiver (or external judge) for judging.

Nevertheless, Hyman and Honorton disagreed over the findings here. Hyman claimed there was a correlation between flaws of randomization and study outcome; Honorton claimed there was not. The sources of this disagreement were in conflicting definitions of flaw categories, in the coding and assignment of flaw ratings to individual studies, and in the subsequent statistical treatment of those ratings.

Unfortunately, there have been no ratings of flaws by independent raters who were unaware of the studies' outcomes (Morris, 1991). Nevertheless, none of the contributors to the subsequent debate concurred with Hyman's conclusion, whereas four nonparapsychologists—two statisticians and two psychologists—explicitly concurred with Honorton's conclusion (Harris & Rosenthal, 1988b; Saunders, 1985; Utts, 1991a). For example, Harris and Rosenthal (one of the pioneers in the use of meta-analysis in psychology) used Hyman's own flaw ratings and failed to find any significant relationships between flaws and study outcomes in each of two separate analyses: "Our analysis of the effects of flaws on study outcome lends no support to the hypothesis that Ganzfeld research results are a significant function of the set of flaw variables" (1988b, p. 3; for a more recent exchange regarding Hyman's analysis, see Hyman, 1991; Utts, 1991a, 1991b).

Effect Size

Some critics of parapsychology have argued that even if current laboratory-produced psi effects turn out to be replicable and nonartifactual, they are too small to be of theoretical interest or practical importance. We do not believe this to be the case for the psi ganzfeld effect.

[2] A 1980 survey of parapsychologists uncovered only 19 completed but unreported ganzfeld studies. Seven of these had achieved significantly positive results, a proportion (.37) very similar to the proportion of independently significant studies in the meta-analysis (.43) (Blackmore, 1980).

In psi ganzfeld studies, the hit rate itself provides a straightforward descriptive measure of effect size, but this measure cannot be compared directly across studies because they do not all use a four-stimulus judging set and, hence, do not all have a chance baseline of .25. The next most obvious candidate, the difference in each study between the hit rate observed and the hit rate expected under the null hypothesis, is also intuitively descriptive but is not appropriate for statistical analysis because not all differences between proportions that are equal are equally detectable (e.g., the power to detect the difference between .55 and .25 is different from the power to detect the difference between .50 and .20).

To provide a scale of equal detectability, Cohen (1988) devised the effect size index h, which involves an arcsine transformation on the proportions before calculation of their difference. Cohen's h is quite general and can assess the difference between any two proportions drawn from independent samples or between a single proportion and any specified hypothetical value. For the 28 studies examined in the meta-analyses, h was .28, with a 95% confidence interval from .11 to .45.

But because values of h do not provide an intuitively descriptive scale, Rosenthal and Rubin (1989; Rosenthal, 1991) have recently suggested a new index, π, which applies specifically to one-sample, multiple-choice data of the kind obtained in ganzfeld experiments. In particular, π expresses all hit rates as the proportion of hits that would have been obtained if there had been only two equally likely alternatives—essentially a coin flip. Thus, π ranges from 0 to 1, with .5 expected under the null hypothesis. The formula is

$$\pi = \frac{P(k-1)}{P(k-2) + 1},$$

where P is the raw proportion of hits and k is the number of alternative choices available. Because π has such a straightforward intuitive interpretation, we use it (or its conversion back to an equivalent four-alternative hit rate) throughout this article whenever it is applicable.

For the 28 studies examined in the meta-analyses, the mean value of π was .62, with a 95% confidence interval from .55 to .69. This corresponds to a four-alternative hit rate of 35%, with a 95% confidence interval from 28% to 43%.

Cohen (1988, 1992) has also categorized effect sizes into *small*, *medium*, and *large*, with *medium* denoting an effect size that should be apparent to the naked eye of a careful observer. For a statistic such as π, which indexes the deviation of a proportion from .5, Cohen considers .65 to be a medium effect size: A statistically unaided observer should be able to detect the bias of a coin that comes up heads on 65% of the trials. Thus, at .62, the psi ganzfeld effect size falls just short of Cohen's naked-eye criterion. From the phenomenology of the ganzfeld experimenter, the corresponding hit rate of 35% implies that he or she will see a subject obtain a hit approximately every third session rather than every fourth.

It is also instructive to compare the psi ganzfeld effect with the results of a recent medical study that sought to determine whether aspirin can prevent heart attacks (Steering Committee of the Physicians' Health Study Research Group, 1988). The study was discontinued after 6 years because it was already clear that the aspirin treatment was effective ($p < .00001$) and it was considered unethical to keep the control group on placebo medication. The study was widely publicized as a major medical breakthrough. But despite its undisputed reality and practical importance, the size of the aspirin effect is quite small: Taking aspirin reduces the probability of suffering a heart attack by only .008. The corresponding effect size (h) is .068, about one third to one fourth the size of the psi ganzfeld effect (Atkinson et al., 1993, p. 236; Utts, 1991b).

In sum, we believe that the psi ganzfeld effect is large enough to be of both theoretical interest and potential practical importance.

Experimental Correlates of the Psi Ganzfeld Effect

We showed earlier that the technique of correlating variables with effect sizes across studies can help to assess whether methodological flaws might have produced artifactual positive outcomes. The same technique can be used more affirmatively to explore whether an effect varies systematically with conceptually relevant variations in experimental procedure. The discovery of such correlates can help to establish an effect as genuine, suggest ways of increasing replication rates and effect sizes, and enhance the chances of moving beyond the simple demonstration of an effect to its explanation. This strategy is only heuristic, however. Any correlates discovered must be considered quite tentative, both because they emerge from post hoc exploration and because they necessarily involve comparisons across heterogeneous studies that differ simultaneously on many interrelated variables, known and unknown. Two such correlates emerged from the meta-analyses of the psi ganzfeld effect.

Single- versus multiple-image targets. Although most of the 28 studies in the meta-analysis used single pictures as targets, 9 (conducted by three different investigators) used *View Master* stereoscopic slide reels that presented multiple images focused on a central theme. Studies using the *View Master* reels produced significantly higher hit rates than did studies using the single-image targets (50% vs. 34%), $t(26) = 2.22$, $p = .035$, two-tailed.

Sender-receiver pairing. In 17 of the 28 studies, participants were free to bring in friends to serve as senders. In 8 studies, only laboratory-assigned senders were used. (Three studies used no sender.) Unfortunately, there is no record of how many participants in the former studies actually brought in friends. Nevertheless, those 17 studies (conducted by six different investigators) had significantly higher hit rates than did the studies that used only laboratory-assigned senders (44% vs. 26%), $t(23) = 2.39$, $p = .025$, two-tailed.

The Joint Communiqué

After their published exchange in 1985, Hyman and Honorton agreed to contribute a joint communiqué to the subsequent discussion that was published in 1986. First they set forth their areas of agreement and disagreement:

> We agree that there is an overall significant effect in this data base that cannot reasonably be explained by selective reporting or multiple analysis. We continue to differ over the degree to which the effect constitutes evidence for psi, but we agree that the final verdict awaits the outcome of future experiments conducted by a broader range of investiga-

tors and according to more stringent standards. (Hyman & Honorton, 1986, p. 351)

They then spelled out in detail the "more stringent standards" they believed should govern future experiments. These standards included strict security precautions against sensory leakage, testing and documentation of randomization methods for selecting targets and sequencing the judging set, statistical correction for multiple analyses, advance specification of the status of the experiment (e.g., pilot study or confirmatory experiment), and full documentation in the published report of the experimental procedures and the status of statistical tests (e.g., planned or post hoc).

The National Research Council Report

In 1988, the National Research Council (NRC) of the National Academy of Sciences released a widely publicized report commissioned by the U.S. Army that assessed several controversial technologies for enhancing human performance, including accelerated learning, neurolinguistic programming, mental practice, biofeedback, and parapsychology (Druckman & Swets, 1988; summarized in Swets & Bjork, 1990). The report's conclusion concerning parapsychology was quite negative: "The Committee finds no scientific justification from research conducted over a period of 130 years for the existence of parapsychological phenomena" (Druckman & Swets, 1988, p. 22).

An extended refutation strongly protesting the committee's treatment of parapsychology has been published elsewhere (Palmer et al., 1989). The pertinent point here is simply that the NRC's evaluation of the ganzfeld studies does not reflect an additional, independent examination of the ganzfeld database but is based on the same meta-analysis conducted by Hyman that we have discussed in this article.

Hyman chaired the NRC's Subcommittee on Parapsychology, and, although he had concurred with Honorton 2 years earlier in their joint communiqué that "there is an overall significant effect in this data base that cannot reasonably be explained by selective reporting or multiple analysis" (p. 351) and that "significant outcomes have been produced by a number of different investigators" (p. 352), neither of these points is acknowledged in the committee's report.

The NRC also solicited a background report from Harris and Rosenthal (1988a), which provided the committee with a comparative methodological analysis of the five controversial areas just listed. Harris and Rosenthal noted that, of these areas, "only the Ganzfeld ESP studies [the only psi studies they evaluated] regularly meet the basic requirements of sound experimental design" (p. 53), and they concluded that

> it would be implausible to entertain the null given the combined *p* from these 28 studies. Given the various problems or flaws pointed out by Hyman and Honorton...we might estimate the obtained accuracy rate to be about 1/3...when the accuracy rate expected under the null is 1/4. (p. 51)[3]

The Autoganzfeld Studies

In 1983, Honorton and his colleagues initiated a new series of ganzfeld studies designed to avoid the methodological problems he and others had identified in earlier studies (Honorton, 1979; Kennedy, 1979). These studies complied with all of the detailed guidelines that he and Hyman were to publish later in their joint communiqué. The program continued until September, 1989, when a loss of funding forced the laboratory to close. The major innovations of the new studies were the computer control of the experimental protocol—hence the name autoganzfeld—and the introduction of videotaped film clips as target stimuli.

Method

The basic design of the autoganzfeld studies was the same as that described earlier[4]: A receiver and sender were sequestered in separate, acoustically-isolated chambers. After a 14-minute period of progressive relaxation, the receiver underwent ganzfeld stimulation while describing his or her thoughts and images aloud for 30 minutes. Meanwhile, the sender concentrated on a randomly selected target. At the end of the ganzfeld period, the receiver was shown four stimuli and, without knowing which of the four had been the target, rated each stimulus for its similarity to his or her mentation during the ganzfeld.

The targets consisted of 80 still pictures (static targets) and 80 short video segments complete with soundtracks (dynamic targets), all recorded on videocassette. The static targets included art prints, photographs, and magazine advertisements; the dynamic targets included excerpts of approximately 1-min duration from motion pictures, TV shows, and cartoons. The 160 targets were arranged in judging sets of four static or four dynamic targets each, constructed to minimize similarities among targets within a set.

Target selection and presentation. The VCR containing the taped targets was interfaced to the controlling computer, which selected the target and controlled its repeated presentation to the sender during the ganzfeld period, thus eliminating the need for a second experimenter to accompany the sender. After the ganzfeld period, the computer randomly sequenced the four-clip judging set and presented it to the receiver on a TV monitor for judging. The receiver used a computer game paddle to make his or her ratings on a 40-point scale that appeared on the TV monitor after each clip was shown. The receiver was permitted to see each clip and to change the ratings repeatedly until he or she was satisfied. The computer then wrote these and other data from the session into a file on a floppy disk. At that point, the sender moved to the receiver's chamber and revealed the identity of the target to both the receiver and the experimenter. Note that the experimenter did not even know the identity of the four-clip judging set until it was displayed to the receiver for judging.

[3]In a troubling development, the chair of the NRC Committee phoned Rosenthal and asked him to delete the parapsychology section of the paper (R. Rosenthal, personal communication, September 15, 1992). Although Rosenthal refused to do so, that section of the Harris-Rosenthal paper is nowhere cited in the NRC report.

[4]Because Honorton and his colleagues have complied with the Hyman-Honorton specification that experimental reports be sufficiently complete to permit others to reconstruct the investigator's procedures, readers who wish to know more detail than we provide here are likely to find whatever they need in the archival publication of these studies in the *Journal of Parapsychology* (Honorton et al., 1990).

Randomization. The random selection of the target and sequencing of the judging set were controlled by a noise-based random number generator interfaced to the computer. Extensive testing confirmed that the generator was providing a uniform distribution of values throughout the full target range (1–160). Tests on the actual frequencies observed during the experiments confirmed that targets were, on average, selected uniformly from among the 4 clips within each target set and that the 4 judging sequences used were uniformly distributed across sessions.

Additional control features. The receiver's and sender's rooms were sound-isolated, electrically shielded chambers with single-door access that could be continuously monitored by the experimenter. There was two-way intercom communication between the experimenter and the receiver but only one-way communication into the sender's room; thus, neither the experimenter nor the receiver could monitor events inside the sender's room. The archival record for each session includes an audiotape containing the receiver's mentation during the ganzfeld period and all verbal exchanges between the experimenter and the receiver throughout the experiment.

The automated ganzfeld protocol has been examined by several dozen parapsychologists and behavioral researchers from other fields, including well-known critics of parapsychology. Many have participated as subjects or observers. All have expressed satisfaction with the handling of security issues and controls.

Parapsychologists have often been urged to employ magicians as consultants to ensure that the experimental protocols are not vulnerable either to inadvertent sensory leakage or to deliberate cheating. Two "mentalists," magicians who specialize in the simulation of psi, have examined the autoganzfeld system and protocol. Ford Kross, a professional mentalist and officer of the mentalist's professional organization, the Psychic Entertainers Association, provided the following written statement "In my professional capacity as a mentalist, I have reviewed Psychophysical Research Laboratories' automated ganzfeld system and found it to provide excellent security against deception by subjects" (personal communication, May, 1989).

Daryl J. Bem has also performed as a mentalist for many years and is a member of the Psychic Entertainers Association. As mentioned in the author note, this article had its origins in a 1983 visit he made to Honorton's laboratory, where he was asked to critically examine the research protocol from the perspective of a mentalist, a research psychologist, and a subject. Needless to say, this article would not exist if he did not concur with Ford Kross's assessment of the security procedures.

Experimental Studies[5]

Altogether, 100 men and 140 women participated as receivers in 354 sessions during the research program. The participants ranged in age from 17 to 74 years ($\overline{m}$ = 37.3, *SD* = 11.8), with a mean formal education of 15.6 years (*SD* = 2.0). Eight separate experimenters, including Honorton, conducted the studies.

The experimental program included three pilot and eight formal studies. Five of the formal studies used novice (first-time) participants who served as the receiver in one session each. The remaining three formal studies used experienced participants.

Pilot studies. Sample sizes were not preset in the three pilot studies. Study 1 comprised 22 sessions and was conducted during the initial development and testing of the autoganzfeld system. Study 2 comprised 9 sessions testing a procedure in which the experimenter, rather than the receiver, served as the judge at the end of the session. Study 3 comprised 35 sessions and served as practice for participants who had completed the allotted number of sessions in the ongoing formal studies but who wanted additional ganzfeld experience. This study also included several demonstration sessions when TV film crews were present.

Novice Studies. Studies 101–104 were each designed to test 50 participants who had had no prior ganzfeld experience; each participant served as the receiver in a single ganzfeld session. Study 104 included 16 of 20 students recruited from the Juilliard School in New York City to test an artistically gifted sample. Study 105 was initiated to accommodate the overflow of participants who had been recruited for Study 104, including the four remaining Juilliard students. The sample size for this study was set to 25, but only 6 sessions had been completed when the laboratory closed. For purposes of exposition, we divided the 56 sessions from Studies 104 and 105 into two parts: Study 104/105(a) comprises the 36 non-Juilliard participants and Study 104/105(b) comprises the 20 Juilliard students.

Study 201. This study was designed to retest the most promising participants from the previous studies. The number of trials was set to 20, but only 7 sessions with 3 participants had been completed when the laboratory closed.

Study 301. This study was designed to compare static and dynamic targets. The sample size was set to 50 sessions. Twenty-five experienced participants each served as the receiver in 2 sessions. Unknown to the participants, the computer control program was modified to ensure that they would each have 1 session with a static target and 1 session with a dynamic target.

Study 302. This study was designed to examine a dynamic target set that had yielded a particularly high hit rate in the previous studies. The study involved experienced participants who had had no prior experience with this particular target set and who were unaware that only one target set was being sampled. Each served as the receiver in a single session. The design called for the study to continue until 15 sessions were completed with each of the targets, but only 25 sessions had been completed when the laboratory closed.

The 11 studies just described comprise all sessions conducted during the 6.5 years of the program. There is no "file drawer" of unreported sessions.

Results

Overall hit rate. As in the earlier meta-analysis, receivers' ratings were analyzed by tallying the proportion of hits achieved and calculating the exact binomial probability for the observed number of hits compared with the chance expectation of .25. As noted earlier, 240 partici-

[5] A recent review of the original computer files uncovered a duplicate record in the autoganzfeld database. This has now been eliminated, reducing by one the number of subjects and sessions. As a result, some of the numbers presented in this article differ slightly from those in Honorton et al. (1990).

Table 1
Outcome by Study

Study	Study/subject description	*N* subjects	*N* trials	*N* hits	% hits	Effect size π	z
1	Pilot	19	22	8	36	.62	0.99
2	Pilot	4	9	3	33	.60	0.25
3	Pilot	24	35	10	29	.55	.032
101	Novice	50	50	12	24	.47	−0.30
102	Novice	50	50	18	36	.63	1.60
103	Novice	50	50	15	30	.55	.067
104/105(a)	Novice	36	36	12	33	.60	0.97
104/105(b)	Juilliard sample	20	20	10	50	.75	2.20
201	Experienced	3	7	3	43	.69	0.69
301	Experienced	25	50	15	30	.56	0.67
302	Experienced	25	25	16	54[a]	.78[a]	3.04[a]
Overall (Studies 1–301)		240	329	106	32	.59	2.89

Note. All z scores are based on the exact binomial probability, with $p = .25$ and $q = .75$.
[a]Adjusted for response bias. The hit rate actually observed was 64%.

pants contributed 354 sessions. For reasons discussed later, Study 302 is analyzed separately, reducing the number of sessions in the primary analysis to 329.

As Table 1 shows, there were 106 hits in the 329 sessions, a hit rate of 32% ($z = 2.89$, $p = .002$, one-tailed), with a 95% confidence interval from 30% to 35%. This corresponds to an effect size (π) of .59, with a 95% confidence interval from .53 to .64.

Table 1 also shows that when Studies 104 and 105 are combined and re-divided into Studies 104/105(a) and 104/105(b), 9 of the 10 studies yield positive effect sizes, with a mean effect size (π) of .61, $t(9) = 4.44$, $p = .0008$ one-tailed. This effect size is equivalent to a four-alternative hit rate of 34%. Alternatively, if Studies 104 and 105 are retained as separate studies, 9 of the 10 studies again yield positive effect sizes, with a mean effect size (π) of .62, $t(9) = 3.73$, $p = .002$, one-tailed. This effect size is equivalent to a four-alternative hit rate of 35% and is identical to that found across the 28 studies of the earlier meta-analysis.[6]

Considered together, sessions with novice participants (Studies 101–105) yielded a statistically significant hit rate of 32.5% ($p = .009$), which is not significantly different from the 31.6% hit rate achieved by experienced participants in Studies 201 and 301. And finally, each of the eight experimenters also achieved a positive effect size, with a mean π of .60, $t(7) = 3.44$, $p = .005$, one-tailed.

The Juilliard sample. There are several reports in the literature of a relationship between creativity or artistic ability and psi performance (Schmeidler, 1988). To explore this possibility in the ganzfeld setting, 10 male and 10 female undergraduates were recruited from the Juilliard School. Of these, 8 were music students, 10 were drama students, and 2 were dance students. Each served as the receiver in a single session in Study 104 or 105. As shown in Table 1, these students achieved a hit rate of 50% ($p = .014$), one of the five highest hit rates ever reported for a single sample in a ganzfeld study. The musicians were particularly successful: 6 of the 8 (75%) successfully identified their targets ($p = .004$; further details about this sample and their ganzfeld performance were reported in Schlitz & Honorton, 1992).

Study size and effect size. There is a significant negative correlation across the 10 studies listed in Table 1 between the number of sessions included in a study and the study's effect size (π), $r = -.64$, $t(8) = 2.36$, $p < .05$, two-tailed. This is reminiscent of Hyman's discovery that the smaller studies in the original ganzfeld database were disproportionately likely to report statistically significant results. He interpreted this finding as evidence for a bias against the reporting of small studies that fail to achieve significant results. A similar interpretation cannot be applied to the autoganzfeld studies, however, because there are no unreported sessions.

One reviewer of this article suggested that the negative correlation might reflect a decline effect in which earlier

[6]As noted above, the laboratory was forced to close before three of the formal studies could be completed. If we assume that the remaining trials in Studies 105 and 201 would have yielded only chance results, this would reduce the overall z for the first 10 autoganzfeld studies from 2.89 to 2.76 ($p = .003$). Thus, inclusion of the two incomplete studies does not pose an optional stopping problem. The third incomplete study, Study 302, is discussed below.

Table 2

Study 302: Expected Hit Rate and Proportion of Sessions in which Each Video Clip was Ranked First when it was a Target and when it was a Decoy

Video Clip	Relative Frequency as Target	Relative Frequency of First Place Ranking	Expected Hit Rate (%)	Ranked First when Target	Ranked First when Decoy	Difference	Fisher's Exact *p*
Tidal Wave	.28 (7/25)	.24 (6/25)	6.72	.57 (4/7)	.11 (2/18)	.46	.032
Snakes	.12 (3/25)	.12 (3/25)	1.44	.67 (2/3)	.05 (1/22)	.62	.029
Sex Scene	.16 (4/25)	.08 (2/25)	1.28	.25 (1/4)	.05 (1/21)	.20	.300
Bugs Bunny	.44 (11/25)	.56 (14/25)	24.64	.82 (9/11)	.36 (5/14)	.46	.027
	Overall		34.08	.58	.14	.44	

sessions of a study are more successful than later sessions. If there were such an effect, then studies with fewer sessions would show larger effect sizes because they would end before a decline could set in. To check this possibility, we computed point-biserial correlations between hits (1) or misses (0) and the session number within each of the 10 studies. All of the correlations hovered around zero; six were positive, four were negative, and the overall mean was .01.

An inspection of Table 1 reveals that the negative correlation derives primarily from the two studies with the largest effect sizes: the 20 sessions with the Juilliard students and the 7 sessions of Study 201, the study specifically designed to retest the most promising participants from the previous studies. Accordingly, it seems likely that the larger effect sizes of these two studies—and hence the significant negative correlation between the number of sessions and the effect size—reflect genuine performance differences between these two small, highly selected samples and other autoganzfeld participants.

Study 302. All of the studies except Study 302 randomly sampled from a pool of 160 static and dynamic targets. Study 302 sampled from a single, dynamic target set that had yielded a particularly high hit rate in the previous studies. The four film clips in this set consisted of a scene of a tidal wave from the movie *Clash of the Titans*, a high-speed sex scene from *A Clockwork Orange*, a scene of crawling snakes from a TV documentary, and a scene from a Bugs Bunny cartoon.

The experimental design called for this study to continue until each of the clips had served as the target 15 times. Unfortunately, the premature termination of this study at 25 sessions left an imbalance in the frequency with which each clip had served as the target. This means that the high hit rate observed (64%) could well be inflated by response biases.

As an illustration, water imagery is frequently reported by receivers in ganzfeld sessions whereas sexual imagery is rarely reported. (Some participants are probably reluctant both to report sexual imagery and to give the highest rating to the sex-related clip.) If a video clip containing popular imagery (such as water) happens to appear as a target more frequently than a clip containing unpopular imagery (such as sex), a high hit rate might simply reflect the coincidence of those frequencies of occurrence with participants' response biases. And, as the second column of Table 2 reveals, the tidal wave clip did in fact appear more frequently as the target than did the sex clip. More generally, the second and third columns of Table 2 show that the frequency with which each film clip was ranked first closely matches the frequency with which each appeared as the target.

One can adjust for this problem by using the observed frequencies in these two columns to compute the hit rate expected if there were no psi effect. In particular, one can multiply each proportion in the second column by the corresponding proportion in the third column—yielding the joint probability that the clip was the target *and* that it was ranked first—and then sum across the four clips. As shown in the fourth column of Table 2, this computation yields an overall expected hit rate of 34.08%. When the observed hit rate of 64% is compared with this baseline, the effect size (h) is .61. As shown in Table 1, this is equivalent to a four-alternative hit rate of 54%, or a π value of .78, and is statistically significant ($z = 3.04$, $p = .0012$).

The psi effect can be seen even more clearly in the remaining columns of Table 2, which control for the differential popularity of the imagery in the clips by displaying how frequently each was ranked first when it was the target compared with how frequently it was ranked first when it was one of the control clips (decoys). As can be seen, each of the four clips was selected as the target relatively more frequently when it was the target than when it was a decoy, a difference that is significant for three of the four clips. On average, a clip was identified as the target 58% of the time when it was the target and only 14% of the time when it was a decoy.

Dynamic versus static targets. The success of Study 302 raises the question of whether dynamic targets are, in general, more effective than static targets. This possibility was also suggested by the earlier meta-analysis, which revealed that studies using multiple-image targets (*View Master* stereoscopic slide reels) obtained significantly higher hit rates than did studies using single-image targets. By adding motion and sound, the video clips might be thought of as high-tech versions of the *View Master* reels.

The 10 autoganzfeld studies that randomly sampled from both dynamic and static target pools yielded 164 sessions with dynamic targets and 165 sessions with static targets.. As predicted, sessions using dynamic targets yielded significantly more hits than did sessions using static targets (37% vs. 27%; Fisher's exact $p < .04$).

Sender-receiver pairing. The earlier meta-analysis revealed that studies in which participants were free to bring in friends to serve as senders produced significantly higher hit rates than studies that used only laboratory-assigned senders. As noted, however, there is no record of how many of the participants in the former studies actually did bring in friends. Whatever the case, sender-receiver pairing was not a significant correlate of psi performance in the autoganzfeld studies: The 197 sessions in which the sender and receiver were friends did not yield a significantly higher proportion of hits than did the 132 sessions in which they were not (35% vs. 29%; Fisher's exact $p = .28$).

Correlations between receiver characteristics and psi performance. Most of the autoganzfeld participants were strong believers in psi: On a 7-point scale, ranging from *strong disbelief in psi* (1) to *strong belief in psi* (7), the mean was 6.2 ($SD = 1.03$); only 2 participants rated their belief in psi below the midpoint of the scale. In addition, 88% of the participants reported personal experiences suggestive of psi, and 80% had some training in meditation or other techniques involving internal focus of attention.

All of these appear to be important variables. The correlation between belief in psi and psi performance is one of the most consistent findings in the parapsychological literature (Palmer, 1978). And within the autoganzfeld studies, successful performance of novice (first-time) participants was significantly predicted by reported personal psi experiences, involvement with meditation or other mental disciplines, and high scores on the Feeling and Perception factors of the Myers-Briggs Type Inventory (Honorton, 1992; Honorton & Schechter, 1987; Myers & McCaulley, 1985). This recipe for success has now been independently replicated in another laboratory (Broughton, Kanthamani, & Khilji, 1990).

The personality trait of extraversion is also associated with better psi performance. A meta-analysis of 60 independent studies with nearly 3,000 subjects revealed a small but reliable positive correlation between extraversion and psi performance, especially in studies that used free-response methods of the kind used in the ganzfeld experiments (Honorton, Ferrari, & Bem, 1992). Across 14 free-response studies conducted by four independent investigators, the correlation for 612 subjects was .20 ($z = 4.82$, $p = 1.5 \times 10^{-6}$). This correlation was replicated in the autoganzfeld studies, in which extraversion scores were available for 218 of the 240 subjects, $r = .18$, $t(216) = 2.67$, $p = .004$, one-tailed.

Finally, there is the strong psi performance of the Juilliard students, discussed earlier, which is consistent with other studies in the parapsychological literature suggesting a relationship between successful psi performance and creativity or artistic ability.

Discussion

Earlier in this article we quoted from the abstract of the Hyman-Honorton communiqué: "We agree that the final verdict awaits the outcome of future experiments conducted by a broader range of investigators and according to more stringent standards" (p. 351). We believe that the "stringent standards" requirement has been met by the autoganzfeld studies. The results are statistically significant and consistent with those in the earlier database. The mean effect size is quite respectable in comparison with other controversial research areas of human performance (Harris & Rosenthal, 1988a). And there are reliable relationships between successful psi performance and conceptually relevant experimental and subject variables, relationships that also replicate previous findings. Hyman (1991) has also commented on the autoganzfeld studies: "Honorton's experiments have produced intriguing results. If...independent laboratories can produce similar results with the same relationships and with the same attention to rigorous methodology, then parapsychology may indeed have finally captured its elusive quarry" (p. 392).

Issues of Replication

As Hyman's comment implies, the autoganzfeld studies by themselves cannot satisfy the requirement that replications be conducted by a "broader range of investigators." Accordingly, we hope the findings reported here will be sufficiently provocative to prompt others to try replicating the psi ganzfeld effect.

We believe that it is essential, however, that future studies comply with the methodological, statistical, and reporting standards set forth in the joint communiqué and achieved by the autoganzfeld studies. It is not necessary for studies to be as automated or as heavily instrumented as the autoganzfeld studies in order to satisfy the methodological guidelines, but they are still likely to be labor intensive and potentially expensive.[7]

Statistical Power and Replication

Would-be replicators also need to be reminded of the power requirements for replicating small effects. Although many academic psychologists do not believe in psi, many apparently do believe in miracles when it comes to replication. Tversky and Kahneman (1971) posed the following problem to their colleagues at meetings of the Mathematical Psychology Group and the American Psychological Association:

> Suppose you have run an experiment on 20 subjects and have obtained a significant result which confirms your the-

[7]As the closing of the autoganzfeld laboratory exemplifies, it is also difficult to obtain funding for psi research. The traditional, peer-refereed sources of funding familiar to psychologists have almost never funded proposals for psi research. The widespread skepticism of psychologists toward psi is almost certainly a contributing factor.

> ory ($z = 2.23$, $p < .05$, two-tailed). You now have cause to run an additional group of 10 subjects. What do you think the probability is that the results will be significant, by a one-tailed test, separately for this group? (p. 105)

The median estimate was .85, with 9 out of 10 respondents providing an estimate greater than .60. The correct answer is approximately .48.

As Rosenthal (1990) has warned: "Given the levels of statistical power at which we normally operate, we have no right to expect the proportion of significant results that we typically do expect, even if in nature there is a very real and very important effect" (p. 16). In this regard, it is again instructive to consider the medical study that found a highly significant effect of aspirin on the incidence of heart attacks. The study monitored more than 22,000 subjects. Had the investigators monitored 3,000 subjects, they would have had less than an even chance of finding a conventionally significant effect. Such is life with small effect sizes.

Given its larger effect size, the prospects for successfully replicating the psi ganzfeld effect are not quite so daunting, but they are probably still grimmer than intuition would suggest. If the true hit rate is in fact about 34% when 25% is expected by chance, then an experiment with 30 trials (the mean for the 28 studies in the original meta-analysis) has only about 1 chance in 6 of finding an effect significant at the .05 level with a one-tailed test. A 50-trial experiment boosts that chance to about 1 in 3. One must escalate to 100 trials in order to come close to the break even point, at which one has a 50–50 chance of finding a statistically significant effect (Utts, 1986). (Recall that only 2 of the 11 autoganzfeld studies yielded results that were individually significant at the conventional .05 level.) Those who require that a psi effect be statistically significant every time before they will seriously entertain the possibility that an effect really exists know not what they ask.

Significance Versus Effect Size

The preceding discussion is unduly pessimistic, however, because it perpetuates the tradition of worshipping the significance level. Regular readers of this journal are likely to be familiar with recent arguments imploring behavioral scientists to overcome their slavish dependence on the significance level as the ultimate measure of virtue and instead to focus more of their attention on effect sizes: "Surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277). Accordingly, we suggest that achieving a respectable effect size with a methodologically tight ganzfeld study would be a perfectly welcome contribution to the replication effort, no matter how untenurable the p level renders the investigator.

Career consequences aside, this suggestion may seem quite counterintuitive. Again, Tversky and Kahneman (1971) have provided an elegant demonstration. They asked several of their colleagues to consider an investigator who runs 15 subjects and obtains a significant t value of 2.46. Another investigator attempts to duplicate the procedure with the same number of subjects and obtains a result in the same direction but with a nonsignificant value of t. Tversky and Kahneman then asked their colleagues to indicate the highest level of t in the replication study they would describe as a failure to replicate. The majority of their colleagues regarded $t = 1.70$ as a failure to replicate. But if the data from two such studies ($t = 2.46$ and $t = 1.70$) were pooled, the t for the combined data would be about 3.00 (assuming equal variances):

> Thus, we are faced with a paradoxical state of affairs, in which the same data that would increase our confidence in the finding when viewed as part of the original study, shake our confidence when viewed as an independent study. (Tversky & Kahneman, 1971, p. 108)

Such is the iron grip of the arbitrary .05. Pooling the data, of course, is what meta-analysis is all about. Accordingly, we suggest that two or more laboratories could collaborate in a ganzfeld replication effort by conducting independent studies and then pooling them in meta-analytic fashion, what one might call real-time meta-analysis. (Each investigator could then claim the pooled p level for his or her own curriculum vitae.)

Maximizing Effect Size

Rather than buying or borrowing larger sample sizes, those who seek to replicate the psi ganzfeld effect might find it more intellectually satisfying to attempt to maximize the effect size by attending to the variables associated with successful outcomes. Thus researchers who wish to enhance the chances of successful replication should use dynamic rather than static targets. Similarly we advise using participants with the characteristics we have reported to be correlated with successful psi performance. Random college sophomores enrolled in introductory psychology do not constitute the optimal subject pool.

Finally, we urge ganzfeld researchers to read carefully the detailed description of the warm social ambiance that Honorton et al. (1990) sought to create in the autoganzfeld laboratory. We believe that the social climate created in psi experiments is a critical determinant of their success or failure.

The Problem of "Other" Variables

This caveat about the social climate of the ganzfeld experiment prompted one reviewer of this article to worry that this provided "an escape clause" that weakens the falsifiability of the psi hypothesis: "Until Bem and Honorton can provide operational criteria for creating a warm social ambiance, the failure of an experiment with otherwise adequate power can always be dismissed as due to a lack of warmth."

Alas, it is true; we devoutly wish it were otherwise. But the operation of unknown variables in moderating the success of replications is a fact of life in all of the sciences. Consider, for example, an earlier article in this journal by Spence (1964). He reviewed studies testing the straightforward derivation from Hullian learning theory that high-anxiety subjects should condition more strongly than low-anxiety subjects. This hypothesis was confirmed 94% of the time in Spence's own laboratory at the University of Iowa but only 63% of the time in laboratories at other universities. In fact, Kimble and his associates at Duke University and the University of North Carolina obtained results in the opposite direction in two of three experiments.

In searching for a post hoc explanation, Spence (1964) noted that "a deliberate attempt was made in the Iowa studies to provide conditions in the laboratory that might elicit some degree of emotionality. Thus, the experimenter was instructed to be impersonal and quite formal ... and did not try to put [subjects] at ease or allay any expressed fears" (pp. 135–136). Moreover, he pointed out, his subjects sat in a dental chair whereas Kimble's subjects sat in a secretarial chair. Spence even considered "the possibility that cultural backgrounds of southern and northern students may lead to a difference in the manner in which they respond to the different items in the [Manifest Anxiety] scale" (p. 136). If this was the state of affairs in an area of research as well established as classical conditioning, then the suggestion that the social climate of the psi laboratory might affect the outcome of ganzfeld experiments in ways not yet completely understood should not be dismissed as a devious attempt to provide an escape clause in case of replication failure.

The best the original researchers can do is to communicate as complete a knowledge of the experimental conditions as possible in an attempt to anticipate some of the relevant moderating variables. Ideally, this might include direct training by the original researchers or videotapes of actual sessions. Lacking these, however, the detailed description of the autoganzfeld procedures provided by Honorton et al. (1990) comes as close as current knowledge permits in providing for other researchers the "operational criteria for creating a warm social ambiance."

Theoretical Considerations

Up to this point, we have confined our discussion to strictly empirical matters. We are sympathetic to the view that one should establish the existence of a phenomenon, anomalous or not, before attempting to explain it. So suppose for the moment that we have a genuine anomaly of information transfer here. How can it be understood or explained?

The Psychology of Psi

In attempting to understand psi, parapsychologists have typically begun with the working assumption that, whatever its underlying mechanisms, it should behave like other, more familiar psychological phenomena. In particular, they typically assume that target information behaves like an external sensory stimulus that is encoded, processed, and experienced in familiar information-processing ways. Similarly, individual psi performances should covary with experimental and subject variables in psychologically sensible ways. These assumptions are embodied in the model of psi that motivated the ganzfeld studies in the first place.

The ganzfeld procedure. As noted in the introduction, the ganzfeld procedure was designed to test a model in which psi-mediated information is conceptualized as a weak signal that is normally masked by internal somatic and external sensory "noise." Accordingly, any technique that raises the signal-to-noise ratio should enhance a person's ability to detect psi-mediated information. This noise-reduction model of psi organizes a large and diverse body of experimental results, particularly those demonstrating the psi-conducive properties of altered states of consciousness such as meditation, hypnosis, dreaming, and, of course, the ganzfeld itself (Rao & Palmer, 1987).

Alternative theories propose that the ganzfeld (and altered states) may be psi-conducive because it lowers resistance to accepting alien imagery, diminishes rational or contextual constraints on the encoding or reporting of information, stimulates more divergent thinking, or even just serves as a placebolike ritual that participants perceive as being psi conducive (Stanford, 1987). At this point, there are no data that would permit one to choose among these alternatives, and the noise-reduction model remains the most widely accepted.

The target. There are also a number of plausible hypotheses that attempt to account for the superiority of dynamic targets over static targets: Dynamic targets contain more information, involve more sensory modalities, evoke more of the receiver's internal schemata, are more lifelike, have a narrative structure, are more emotionally evocative, and are "richer" in other, unspecified ways. Several psi researchers have attempted to go beyond the simple dynamic-static dichotomy to more refined or theory-based definitions of a good target. Although these efforts have involved examining both psychological and physical properties of targets, there is as yet not much progress to report (Delanoy, 1990).

The receiver. Some of the subject characteristics associated with good psi performance also appear to have psychologically straightforward explanations. For example, garden-variety motivational explanations seem sufficient to account for the relatively consistent finding that those who believe in psi perform significantly better than those who do not. (Less straightforward, however, would be an explanation for the frequent finding that nonbelievers actually perform significantly worse than chance [Broughton, 1991, p. 109].)

The superior psi performance of creative or artistically gifted individuals—like the Juilliard students—may reflect individual differences that parallel some of the hypothesized effects of the ganzfeld mentioned earlier: Artistically gifted individuals may be more receptive to alien imagery, be better able to transcend rational or contextual constraints on the encoding or reporting of information, or be more divergent in their thinking. It has also been suggested that both artistic and psi abilities might be rooted in superior right-brain functioning.

The observed relationship between extraversion and psi performance has been of theoretical interest for many years. Eysenck (1966) reasoned that extraverts should perform well in psi tasks because they are easily bored and respond favorably to novel stimuli. In a setting such as the ganzfeld, extraverts may become "stimulus starved" and thus be highly sensitive to any stimulation, including weak incoming psi information. In contrast, introverts would be more inclined to entertain themselves with their own thoughts and thus continue to mask psi information despite the diminished sensory input. Eysenck also speculated that psi might be a primitive form of perception antedating cortical developments in the course of evolution, and, hence, cortical arousal might suppress psi functioning. Because extraverts have a lower level of cortical arousal than introverts, they should perform better in psi tasks (the evolutionary biology of psi has also been discussed by Broughton, 1991, pp. 347–352).

But there are more mundane possibilities. Extraverts might perform better than introverts simply because they are more relaxed and comfortable in the social setting of the typical psi experiment (e.g., the "warm social ambiance" of the autoganzfeld studies). This interpretation is strengthened by the observation that introverts outperformed extraverts in a study in which subjects had no contact with an experimenter but worked alone at home with materials they received in the mail (Schmidt & Schlitz, 1989). To help decide among these interpretations, ganzfeld experimenters have begun to use the extraversion scale of the NEO Personality Inventory (Costa & Mc-

Crae, 1992), which assesses six different facets of the extraversion-introversion factor.

The sender. In contrast to this information about the receiver in psi experiments, virtually nothing is known about the characteristics of a good sender or about the effects of the sender's relationship with the receiver. As has been shown, the initial suggestion from the meta-analysis of the original ganzfeld database that psi performance might be enhanced when the sender and receiver are friends was not replicated at a statistically significant level in the autoganzfeld studies.

A number of parapsychologists have entertained the more radical hypothesis that the sender may not even be a necessary element in the psi process. In the terminology of parapsychology, the sender-receiver procedure tests for the existence of *telepathy*, anomalous communication between two individuals; however if the receiver is somehow picking up the information from the target itself, it would be termed *clairvoyance*, and the presence of the sender would be irrelevant (except for possible psychological reasons such as expectation effects).

At the time of his death, Honorton was planning a series of autoganzfeld studies that would systematically compare sender and no-sender conditions while keeping both the receiver and the experimenter blind to the condition of the ongoing session. In preparation, he conducted a meta-analytic review of ganzfeld studies that used no sender. He found 12 studies with a median of 33.5 sessions, conducted by seven investigators. The overall effect size (π) was .56, which corresponds to a four-alternative hit rate of 29%. But this effect size does not reach statistical significance (Stouffer $z = 1.31$, $p = .095$). So far, then, there is no firm evidence for psi in the ganzfeld in the absence of a sender. (There are, however, nonganzfeld studies in the literature that do report significant evidence for clairvoyance, including a classic card-guessing experiment conducted by J. B. Rhine and Pratt [1954].)

The Physics of Psi

The psychological level of theorizing discussed earlier does not, of course, address the conundrum that makes psi phenomena anomalous in the first place: their presumed incompatibility with our current conceptual model of physical reality. Parapsychologists differ widely from one another in their taste for theorizing at this level, but several whose training lies in physics or engineering have proposed physical (or biophysical) theories of psi phenomena (an extensive review of theoretical parapsychology was provided by Stokes, 1987). Only some of these theories would force a radical revision in our conception of physical reality.

Those who follow contemporary debates in modern physics, however, will be aware that several phenomena predicted by quantum theory and confirmed by experiment are themselves incompatible with our current conceptual model of physical reality. Of these, it is the 1982 empirical confirmation of Bell's theorem that has created the most excitement and controversy among philosophers and the few physicists who are willing to speculate on such matters (Cushing & McMullin, 1989; Herbert, 1987). In brief, Bell's theorem states that any model of reality that is compatible with quantum mechanics must be *nonlocal*: It must allow for the possibility that the results of observations at two arbitrarily distant locations can be correlated in ways that are incompatible with any physically permissible causal mechanism.

Several possible models of reality that incorporate nonlocality have been proposed by both philosophers and physicists. Some of these models clearly rule out psi-like information transfer, others permit it, and some actually require it. Thus, at a grander level of theorizing, some parapsychologists believe that one of the more radical models of reality compatible with both quantum mechanics and psi will eventually come to be accepted. If and when that occurs, psi phenomena would cease to be anomalous.

But we have learned that all such talk provokes most of our colleagues in psychology and in physics to roll their eyes and gnash their teeth. So let's just leave it at that.

Skepticism Revisited

More generally, we have learned that our colleagues' tolerance for *any* kind of theorizing about psi is strongly determined by the degree to which they have been convinced by the data that psi has been demonstrated. We have further learned that their diverse reactions to the data themselves are strongly determined by their a priori beliefs about and attitudes toward a number of quite general issues, some scientific, some not. In fact, several statisticians believe that the traditional hypothesis testing methods used in the behavioral sciences should be abandoned in favor of Bayesian analyses, which take into account a person's a priori beliefs about the phenomenon under investigation (e.g., Bayarri & Berger, 1991; Dawson, 1991).

In the final analysis, however, we suspect that both one's Bayesian a prioris *and* one's reactions to the data are ultimately determined by whether one was more severely punished in childhood for Type I or Type II errors.

References

Atkinson, R., Atkinson, R. C., Smith, E. E., & Bem, D. J. (1990). *Introduction to psychology* (10th ed.). San Diego, CA: Harcourt Brace Jovanovich.

Atkinson, R., Atkinson, R. C., Smith, E. E., & Bem, D. J. (1993). *Introduction to psychology* (11th ed.). San Diego, CA: Harcourt Brace Jovanovich.

Avant, L. L. (1965). Vision in the ganzfeld. *Psychological Bulletin, 64*, 246-258.

Bayarri, M. J., & Berger, J. (1991). Comment. *Statistical Science, 6*, 379-382.

Blackmore, S. (1980). The extent of selective reporting of ESP Ganzfeld studies. *European Journal of Parapsychology, 3*, 213-219.

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist, 27*, 774-775.

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.

Broughton, R. S. (1991). *Parapsychology: The controversial science*. New York: Ballantine Books.

Broughton, R. S., Kanthamani, H., & Khilji, A. (1990). Assessing the PRL success model on an independent ganzfeld data base. In L. Henkel & J. Palmer (Eds.), *Research in parapsychology 1989* (pp. 32-35). Metuchen, NJ: Scarecrow Press.

Child, I. L. (1985). Psychology and anomalous observations: The question of ESP in dreams. *American Psychologist, 40,* 1219-1230.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1,* 98-101.

Costa, P. T. J., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Manual.* Odessa, FL: Psychological Assessment Resources.

Cushing, J. T., & McMullin, E. (Eds.). (1989). *Philosophical consequences of quantum theory: Reflections on Bell's theorem.* Notre Dame, IN: University of Notre Dame Press.

Dawson, R. (1991). Comment. *Statistical Science, 6,* 382-385.

Delanoy, D. L. (1990). Approaches to the target: A time for reevaluation. In L. A. Henkel, & J. Palmer (Eds.), *Research in Parapsychology 1989* (pp. 89-92). Metuchen, NJ: Scarecrow Press.

Dingwall, E. J. (Ed.). (1968). *Abnormal hypnotic phenomena* (4 vols.). London: Churchill.

Druckman, D., & Swets, J. A. (Eds.). (1988). *Enhancing human performance. Issues, theories, and techniques.* Washington, DC: National Academy Press.

Eysenck, H. J. (1966). Personality and extra-sensory perception. *Journal of the Society for Psychical Research, 44,* 55-71.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York: Free Press.

Green, C. E. (1960). Analysis of spontaneous cases. *Proceedings of the Society for Psychical Research, 53,* 97-161.

Harris, M. J., & Rosenthal, R. (1988a). *Human performance research: An overview.* Washington, DC: National Academy Press.

Harris, M. J., & Rosenthal, R. (1988b). Postscript to "Human performance research: An overview." Washington, DC: National Academy Press.

Herbert, N. (1987). *Quantum reality: Beyond the new physics.* Garden City, NY: Anchor Books.

Honorton, C. (1969). Relationship between EEG alpha activity and ESP card-guessing performance. *Journal of the American Society for Psychical Research, 63,* 365-374.

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. (1979). Methodological issues in free-response experiments. *Journal of the American Society for Psychical Research, 73,* 381-394.

Honorton, C. (1985). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology, 49,* 51-91.

Honorton, C. (1992). The ganzfeld novice: Four predictors of initial ESP performance. *Proceedings of the Parapsychological Association 35th Annual Convention, Las Vegas, NV,* 51-58.

Honorton, C., Berger, R. E., Varvoglis, M. P., Quant, M., Derr, P., Schechter, E. I., & Ferrari, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology, 54,* 99-139.

Honorton, C., Ferrari, D. C., & Bem, D. J. (1992). Extraversion and ESP performance: Meta-analysis and a new confirmation. In L. A. Henkel & G. R. Schmeidler (Eds.), *Research in parapsychology 1990* (pp. 35-38). Metuchen, NJ: Scarecrow Press.

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., & Schechter, E. I. (1987). Ganzfeld target retrieval with an automated testing system: A model for initial ganzfeld success. In D. B. Weiner & R. D. Nelson (Eds.), *Research in parapsychology 1986* (pp. 36-39). Metuchen, NJ: Scarecrow Press.

Hyman, R. (1985). The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology, 49,* 3-49.

Hyman, R. (1991). Comment. *Statistical Science, 6,* 389-392.

Hyman, R., & Honorton, C. (1986). A joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology, 50,* 351-364.

Kennedy, J. E. (1979). Methodological problems in free-response ESP experiments. *Journal of the American Society for Psychical Research, 73,* 1-15.

Metzger, W. (1930). Optische Untersuchungen am Ganzfeld: II. Zur phanomenologie des homogenen Ganzfelds [Optical investigation of the Ganzfeld: II Toward the phenomenology of the homogeneous Ganzfeld]. *Psychologische Forschung, 13,* 6-29.

Morris, R. L. (1991). Comment. *Statistical Science, 6,* 393-395.

Myers, I. B., & McCaulley, M. H. (1985). *Manual: A guide to the development and use of the Myers-Briggs Type Indicator.* Palo Alto, CA: Consulting Psychologists Press.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Palmer, J. (1978). Extrasensory perception: Research findings. In S. Krippner (Ed.), *Advances in parapsychological research* (Vol. 2, pp. 59-243). New York: Plenum.

Palmer, J. A., Honorton, C., & Utts, J. (1989). Reply to the National Research Council Study on Parapsychology. *Journal of the American Society for Psychical Research, 83,* 31-49.

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 Press.

Parker, A. (1978). A holistic methodology in psi research. *Parapsychology Review, 9,* 1-6.

Prasad, J., & Stevenson, I. (1968). A survey of spontaneous psychical experiences in school children of Uttar Pradesh, India. *International Journal of Parapsychology, 10,* 241-261.

Rao, K. R., & Palmer, J. (1987). The anomaly called psi: Recent research and criticism. *Behavioral and Brain Sciences, 10,* 539-551.

Rhine, J. B., & Pratt, J. G. (1954). A review of the Pearce-Pratt distance series of ESP tests. *Journal of Parapsychology, 18,* 165-177.

Rhine, L. E. (1962). Psychological processes in ESP experiences. I. Waking experiences. *Journal of Parapsychology, 26,* 88-111.

Roig, M., Icochea, H., & Cuzzucoli, A. (1991). Coverage of parapsychology in introductory psychology textbooks. *Teaching of Psychology, 18*, 157-160.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85*, 185-193.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality, 5*, 1-30.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R., & Rubin, D. B. (1989). Effect size estimation for one-sample multiple-choice-type data: Design, analysis, and meta-analysis. *Psychological Bulletin, 106*, 332-337.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Sannwald, G. (1959). Statistische untersuchungen an Spontanphänomene [Statistical investigation of spontaneous phenomena]. *Zeitschrift für Parapsychologie and Grenzgebiete der Psychologie, 3*, 59-71.

Saunders, D. R. (1985). On Hyman's factor analyses. *Journal of Parapsychology, 49*, 86-88.

Schechter, E. I. (1984). Hypnotic induction vs. control conditions: Illustrating an approach to the evaluation of replicability in parapsychology. *Journal of the American Society for Psychical Research, 78*, 1-27.

Schlitz, M. J., & Honorton, C. (1992). Ganzfeld psi performance within an artistically gifted population. *Journal of the American Society for Psychical Research, 86*, 83-98.

Schmeidler, G. R. (1988). *Parapsychology and psychology: Matches and Mismatches*. Jefferson, NC: McFarland.

Schmidt, H., & Schlitz, M. J. (1989). A large scale pilot PK experiment with prerecorded random events. In L. A. Henkel & R. E. Berger (Eds.), *Research in Parapsychology 1988* (pp. 6-10). Metuchen, NJ: Scarecrow Press.

Spence, K. W. (1964). Anxiety (drive) level and performance in eyelid conditioning. *Psychological Bulletin, 61*, 129-139.

Stanford, R. G. (1987). Ganzfeld and hypnotic-induction procedures in ESP research: Toward understanding their success. In S. Krippner (Ed.), *Advances in parapsychological research* (Vol. 5, pp. 39-76). Jefferson, NC: McFarland.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine, 318*, 262-264.

Sterling, T. C. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association, 54*, 30-34.

Stokes, D. M. (1987). Theoretical parapsychology. In S. Krippner (Ed.), *Advances in parapsychological research* (Vol. 5, pp. 77-189). Jefferson, NC: McFarland.

Swets, J. A., & Bjork, R. A. (1990). Enhancing human performance: An evaluation of "new age" techniques considered by the U. S. Army. *Psychological Science, 1*, 85-96.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 2*, 105-110.

Ullman, M., Krippner, S., & Vaughan, A. (1973). *Dream telepathy*. New York: Macmillan.

Utts, J. (1986). The ganzfeld debate: A statistician's perspective. *Journal of Parapsychology, 50*, 393-402.

Utts, J. (1991a). Rejoinder. *Statistical Science, 6*, 396-403.

Utts, J. (1991b). Replication and meta-analysis in parapsychology. *Statistical Science, 6*, 363-378.

Wagner, M. W., & Monnet, M. (1979). Attitudes of college professors toward extra-sensory perception. *Zetetic Scholar, 5*, 7-17.

Received September 28, 1992
Revision received March 10, 1993
Accepted March 14, 1993

Statistical Science
1991, Vol. 6, No. 4, 363–403

Replication and Meta-Analysis in Parapsychology

Jessica Utts

Abstract. Parapsychology, the laboratory study of psychic phenomena, has had its history interwoven with that of statistics. Many of the controversies in parapsychology have focused on statistical issues, and statistical models have played an integral role in the experimental work. Recently, parapsychologists have been using meta-analysis as a tool for synthesizing large bodies of work. This paper presents an overview of the use of statistics in parapsychology and offers a summary of the meta-analyses that have been conducted. It begins with some anecdotal information about the involvement of statistics and statisticians with the early history of parapsychology. Next, it is argued that most nonstatisticians do not appreciate the connection between power and "successful" replication of experimental effects. Returning to parapsychology, a particular experimental regime is examined by summarizing an extended debate over the interpretation of the results. A new set of experiments designed to resolve the debate is then reviewed. Finally, meta-analyses from several areas of parapsychology are summarized. It is concluded that the overall evidence indicates that there is an anomalous effect in need of an explanation.

Key words and phrases: Effect size, psychic research, statistical controversies, randomness, vote-counting.

1. INTRODUCTION

In a June 1990 Gallup Poll, 49% of the 1236 respondents claimed to believe in extrasensory perception (ESP), and one in four claimed to have had a personal experience involving telepathy (Gallup and Newport, 1991). Other surveys have shown even higher percentages; the University of Chicago's National Opinion Research Center recently surveyed 1473 adults, of which 67% claimed that they had experienced ESP (Greeley, 1987).

Public opinion is a poor arbiter of science, however, and experience is a poor substitute for the scientific method. For more than a century, small numbers of scientists have been conducting laboratory experiments to study phenomena such as telepathy, clairvoyance and precognition, collectively known as "psi" abilities. This paper will examine some of that work, as well as some of the statistical controversies it has generated.

Parapsychology, as this field is called, has been a source of controversy throughout its history. Strong beliefs tend to be resistant to change even in the face of data, and many people, scientists included, seem to have made up their minds on the question without examining any empirical data at all. A critic of parapsychology recently acknowledged that "The level of the debate during the past 130 years has been an embarrassment for anyone who would like to believe that scholars and scientists adhere to standards of rationality and fair play" (Hyman, 1985a, page 89). While much of the controversy has focused on poor experimental design and potential fraud, there have been attacks and defenses of the statistical methods as well, sometimes calling into question the very foundations of probability and statistical inference.

Most of the criticisms have been leveled by psychologists. For example, a 1988 report of the U.S. National Academy of Sciences concluded that "The committee finds no scientific justification from research conducted over a period of 130 years for the existence of parapsychological phenomena" (Druckman and Swets, 1988, page 22). The chapter on parapsychology was written by a subcommittee

Jessica Utts is Associate Professor, Division of Statistics, University of California at Davis, 469 Kerr Hall, Davis, California 95616.

chaired by a psychologist who had published a similar conclusion prior to his appointment to the committee (Hyman, 1985a, page 7). There were no parapsychologists involved with the writing of the report. Resulting accusations of bias (Palmer, Honorton and Utts, 1989) led U.S. Senator Claiborne Pell to request that the Congressional Office of Technology Assessment (OTA) conduct an investigation with a more balanced group. A one-day workshop was held on September 30, 1988, bringing together parapsychologists, critics and experts in some related fields (including the author of this paper). The report concluded that parapsychology needs "a fairer hearing across a broader spectrum of the scientific community, so that emotionality does not impede objective assessment of experimental results" (Office of Technology Assessment, 1989).

It is in the spirit of the OTA report that this article is written. After Section 2, which offers an anecdotal account of the role of statisticians and statistics in parapsychology, the discussion turns to the more general question of replication of experimental results. Section 3 illustrates how replication has been (mis)interpreted by scientists in many fields. Returning to parapsychology in Section 4, a particular experimental regime called the "ganzfeld" is described, and an extended debate about the interpretation of the experimental results is discussed. Section 5 examines a meta-analysis of recent ganzfeld experiments designed to resolve the debate. Finally, Section 6 contains a brief account of meta-analyses that have been conducted in other areas of parapsychology, and conclusions are given in Section 7.

2. STATISTICS AND PARAPSYCHOLOGY

Parapsychology had its beginnings in the investigation of purported mediums and other anecdotal claims in the late 19th century. The Society for Psychical Research was founded in Britain in 1882, and its American counterpart was founded in Boston in 1884. While these organizations and their members were primarily involved with investigating anecdotal material, a few of the early researchers were already conducting "forced-choice" experiments such as card-guessing. (Forced-choice experiments are like multiple choice tests; on each trial the subject must guess from a small, known set of possibilities.) Notable among these was Nobel Laureate Charles Richet, who is generally credited with being the first to recognize that probability theory could be applied to card-guessing experiments (Rhine, 1977, page 26; Richet, 1884).

F. Y. Edgeworth, partly in response to what he considered to be incorrect analyses of these experiments, offered one of the earliest treatises on the statistical evaluation of forced-choice experiments in two articles published in the *Proceedings of the Society for Psychical Research* (Edgeworth, 1885, 1886). Unfortunately, as noted by Mauskopf and McVaugh (1979) in their historical account of the period, Edgeworth's papers were "perhaps too difficult for their immediate audience" (page 105).

Edgeworth began his analysis by using Bayes' theorem to derive the formula for the posterior probability that chance was operating, given the data. He then continued with an argument "savouring more of Bernoulli than Bayes" in which "it is consonant, I submit, to experience, to put 1/2 both for α and β," that is, for both the prior probability that chance alone was operating, and the prior probability that "there should have been some additional agency." He then reasoned (using a Taylor series expansion of the posterior probability formula) that if there were a large probability of observing the data given that some additional agency was at work, and a small objective probability of the data under chance, then the latter (binomial) probability "may be taken as a rough measure of the sought *a posteriori* probability in favour of mere chance" (page 195). Edgeworth concluded his article by applying his method to some data published previously in the same journal. He found the probability against chance to be 0.99996, which he said "may fairly be regarded as physical certainty" (page 199). He concluded:

> Such is the evidence which the calculus of probabilities affords as to the existence of an agency other than mere chance. The calculus is silent as to the nature of that agency—whether it is more likely to be vulgar illusion or extraordinary law. That is a question to be decided, not by formulae and figures, but by general philosophy and common sense [page 199].

Both the statistical arguments and the experimental controls in these early experiments were somewhat loose. For example, Edgeworth treated as binomial an experiment in which one person chose a string of eight letters and another attempted to guess the string. Since it has long been understood that people are poor random number (or letter) generators, there is no statistical basis for analyzing such an experiment. Nonetheless, Edgeworth and his contemporaries set the stage for the use of controlled experiments with statistical evaluation in laboratory parapsychology. An interesting historical account of Edgeworth's involvement and the role telepathy experiments played in the early history of randomization and experimental design is provided by Hacking (1988).

One of the first American researchers to use statistical methods in parapsychology was John Edgar Coover, who was the Thomas Welton Stanford Psychical Research Fellow in the Psychology Department at Stanford University from 1912 to 1937 (Dommeyer, 1975). In 1917, Coover published a large volume summarizing his work (Coover, 1917). Coover believed that his results were consistent with chance, but others have argued that Coover's definition of significance was too strict (Dommeyer, 1975). For example, in one evaluation of his telepathy experiments, Coover found a two-tailed p-value of 0.0062. He concluded, "Since this value, then, lies within the field of chance deviation, although the probability of its occurrence by chance is fairly low, it cannot be accepted as a decisive indication of some cause beyond chance which operated in favor of success in guessing" (Coover, 1917, page 82). On the next page, he made it explicit that he would require a p-value of 0.0000221 to declare that something other than chance was operating.

It was during the summer of 1930, with the card-guessing experiments of J. B. Rhine at Duke University, that parapsychology began to take hold as a laboratory science. Rhine's laboratory still exists under the name of the Foundation for Research on the Nature of Man, housed at the edge of the Duke University campus.

It wasn't long after Rhine published his first book, *Extrasensory Perception* in 1934, that the attacks on his methodology began. Since his claims were wholly based on statistical analyses of his experiments, the statistical methods were closely scrutinized by critics anxious to find a conventional explanation for Rhine's positive results.

The most persistent critic was a psychologist from McGill University named Chester Kellogg (Mauskopf and McVaugh, 1979). Kellogg's main argument was that Rhine was using the binomial distribution (and normal approximation) on a series of trials that were not independent. The experiments in question consisted of having a subject guess the order of a deck of 25 cards, with five each of five symbols, so technically Kellogg was correct.

By 1937, several mathematicians and statisticians had come to Rhine's aid. Mauskopf and McVaugh (1979) speculated that since statistics was itself a young discipline, "a number of statisticians were equally outraged by Kellogg, whose arguments they saw as discrediting *their* profession" (page 258). The major technical work, which acknowledged that Kellogg's criticisms were accurate but did little to change the significance of the results, was conducted by Charles Stuart and Joseph A. Greenwood and published in the first [illegible] and Greenwood, 1937). Stuart, who had been an undergraduate in mathematics at Duke, was one of Rhine's early subjects and continued to work with him as a researcher until Stuart's death in 1947. Greenwood was a Duke mathematician, who apparently converted to a statistician at the urging of Rhine.

Another prominent figure who was distressed with Kellogg's attack was E. V. Huntington, a mathematician at Harvard. After corresponding with Rhine, Huntington decided that, rather than further confuse the public with a technical reply to Kellogg's arguments, a simple statement should be made to the effect that the mathematical issues in Rhine's work had been resolved. Huntington must have successfully convinced his former student, Burton Camp of Wesleyan, that this was a wise approach. Camp was the 1937 President of IMS. When the annual meetings were held in December of 1937 (jointly with AMS and AAAS), Camp released a statement to the press that read:

> Dr. Rhine's investigations have two aspects: experimental and statistical. On the experimental side mathematicians, of course, have nothing to say. On the statistical side, however, recent mathematical work has established the fact that, assuming that the experiments have been properly performed, the statistical analysis is essentially valid. If the Rhine investigation is to be fairly attacked, it must be on other than mathematical grounds [Camp, 1937].

One statistician who did emerge as a critic was William Feller. In a talk at the Duke Mathematical Seminar on April 24, 1940, Feller raised three criticisms to Rhine's work (Feller, 1940). They had been raised before by others (and continue to be raised even today). The first was that inadequate shuffling of the cards resulted in additional information from one series to the next. The second was what is now known as the "file-drawer effect," namely, that if one combines the results of published studies only, there is sure to be a bias in favor of successful studies. The third was that the results were enhanced by the use of optional stopping, that is, by not specifying the number of trials in advance. All three of these criticisms were addressed in a rejoinder by Greenwood and Stuart (1940), but Feller was never convinced. Even in its third edition published in 1968, his book *An Introduction to Probability Theory and Its Applications* still contains his conclusion about Greenwood and Stuart: "Both their arithmetic and their experiments have a distinct tinge of the supernatural" (Feller, 1968, page 407). In his discussion of Feller's [illegible] "I believe

Feller was confused... he seemed to have decided the opposition was wrong and that was that."

Several statisticians have contributed to the literature in parapsychology to greater or lesser degrees. T. N. E. Greville developed applicable statistical methods for many of the experiments in parapsychology and was Statistical Editor of the *Journal of Parapsychology* (with J. A. Greenwood) from its start in 1937 through Volume 31 in 1967; Fisher (1924, 1929) addressed some specific problems in card-guessing experiments; Wilks (1965a, b) described various statistical methods for parapsychology; Lindley (1957) presented a Bayesian analysis of some parapsychology data; and Diaconis (1978) pointed out some problems with certain experiments and presented a method for analyzing experiments when feedback is given.

Occasionally, attacks on parapsychology have taken the form of attacks on statistical inference in general, at least as it is applied to real data. Spencer-Brown (1957) attempted to show that true randomness is impossible, at least in finite sequences, and that this could be the explanation for the results in parapsychology. That argument reemerged in a recent debate on the role of randomness in parapsychology, initiated by psychologist J. Barnard Gilmore (Gilmore, 1989, 1990; Utts, 1989; Palmer, 1989, 1990). Gilmore stated that "The agnostic statistician, advising on research in psi, should take account of the possible inappropriateness of classical inferential statistics" (1989, page 338). In his second paper, Gilmore reviewed several non-psi studies showing purportedly random systems that do not behave as they should under randomness (e.g., Iversen, Longcor, Mosteller, Gilbert and Youtz, 1971; Spencer-Brown, 1957). Gilmore concluded that "Anomalous data... should not be found nearly so often if classical statistics offers a valid model of reality" (1990, page 54), thus rejecting the use of classical statistical inference for real-world applications in general.

3. REPLICATION

Implicit and explicit in the literature on parapsychology is the assumption that, in order to truly establish itself, the field needs to find a repeatable experiment. For example, Diaconis (1978) started the summary of his article in *Science* with the words "In search of repeatable ESP experiments, modern investigators..." (page 131). On October 28–29, 1983, the 32nd International Conference of the Parapsychology Foundation was held in San Antonio, Texas, to address "The Repeatability Problem in Parapsychology." The Conference Proceedings (Shapin and Coly, 1985) reflect the diverse views among parapsychologists on the nature of the problem. Honorton (1985a) and Rao (1985), for example, both argued that strict replication is uncommon in *most* branches of science and that parapsychology should not be singled out as unique in this regard. Other authors expressed disappointment in the lack of a single repeatable experiment in parapsychology, with titles such as "Unrepeatability: Parapsychology's Only Finding" (Blackmore, 1985), and "Research Strategies for Dealing with Unstable Phenomena" (Beloff, 1985).

It has never been clear, however, just exactly what would constitute acceptable evidence of a repeatable experiment. In the early days of investigation, the major critics "insisted that it would be sufficient for Rhine and Soal to convince them of ESP if a parapsychologist could perform successfully a single 'fraud-proof' experiment" (Hyman, 1985a, page 71). However, as soon as well-designed experiments showing statistical significance emerged, the critics realized that a single experiment could be statistically significant just by chance. British psychologist C. E. M. Hansel quantified the new expectation, that the experiment should be repeated a few times, as follows:

> If a result is significant at the .01 level and this result is not due to chance but to information reaching the subject, it may be expected that by making two further sets of trials the antichance odds of one hundred to one will be increased to around a million to one, thus enabling the effects of ESP—or whatever is responsible for the original result—to manifest itself to such an extent that there will be little doubt that the result is not due to chance [Hansel, 1980, page 298].

In other words, three consecutive experiments at $p \leq 0.01$ would convince Hansel that something other than chance was at work.

This argument implies that if a particular experiment produces a statistically significant result, but subsequent replications fail to attain significance, then the original result was probably due to chance, or at least remains unconvincing. The problem with this line of reasoning is that there is no consideration given to sample size or power. Only an experiment with extremely high power should be expected to be "successful" three times in succession.

It is perhaps a failure of the way statistics is taught that many scientists do not understand the importance of power in defining successful replication. To illustrate this point, psychologists Tversky and Kahnemann (1982) distributed a questionnaire to their colleagues at a professional meeting, with the question:

> An investigator has reported a result that you consider implausible. He ran 15 subjects, and reported a significant value, $t = 2.46$. Another investigator has attempted to duplicate his procedure, and he obtained a nonsignificant value of t with the same number of subjects. The direction was the same in both sets of data. You are reviewing the literature. What is the highest value of t in the second set of data that you would describe as a failure to replicate? [1982, page 28].

In reporting their results, Tversky and Kahnemann stated:

> The majority of our respondents regarded $t = 1.70$ as a failure to replicate. If the data of two such studies ($t = 2.46$ and $t = 1.70$) are pooled, the value of t for the combined data is about 3.00 (assuming equal variances). Thus, we are faced with a paradoxical state of affairs, in which the same data that would increase our confidence in the finding when viewed as part of the original study, shake our confidence when viewed as an independent study [1982, page 28].

At a recent presentation to the History and Philosophy of Science Seminar at the University of California at Davis, I asked the following question. Two scientists, Professors A and B, each have a theory they would like to demonstrate. Each plans to run a fixed number of Bernoulli trials and then test H_0: $p = 0.25$ versus H_a: $p > 0.25$. Professor A has access to large numbers of students each semester to use as subjects. In his first experiment, he runs 100 subjects, and there are 33 successes ($p = 0.04$, one-tailed). Knowing the importance of replication, Professor A runs an additional 100 subjects as a second experiment. He finds 36 successes ($p = 0.009$, one-tailed).

Professor B only teaches small classes. Each quarter, she runs an experiment on her students to test her theory. She carries out ten studies this way, with the results in Table 1.

I asked the audience by a show of hands to indicate whether or not they felt the scientists had successfully demonstrated their theories. Professor A's theory received overwhelming support, with approximately 20 votes, while Professor B's theory received only one vote.

If you aggregate the results of the experiments for each professor, you will notice that each conducted 200 trials, and Professor B actually demonstrated a *higher* level of success than Professor A, with 71 as opposed to 69 successful trials. The one-tailed p-values for the combined trials are 0.0017 for Professor A and 0.0006 for Professor B.

To address the question of replication more explicitly, I also posed the following scenario. In December of 1987, it was decided to prematurely terminate a study on the effects of aspirin in reducing heart attacks because the data were so convincing (see, e.g., Greenhouse and Greenhouse, 1988; Rosenthal, 1990a). The physician-subjects had been randomly assigned to take aspirin or a placebo. There were 104 heart attacks among the 11,037 subjects in the aspirin group, and 189 heart attacks among the 11,034 subjects in the placebo group (chi-square = 25.01, $p < 0.00001$).

After showing the results of that study, I presented the audience with two hypothetical experiments conducted to try to replicate the original result, with outcomes in Table 2.

I asked the audience to indicate which one they thought was a more successful replication. The audience chose the second one, as would most journal editors, because of the "significant p-value." In fact, the *first* replication has almost exactly the same proportion of heart attacks in the two groups as the original study and is thus a very close replication of that result. The second replication has

TABLE 1
Attempted replciations for professor B

n	Number of successes	One-tailed p-value
10	4	0.22
15	6	0.15
17	6	0.23
25	8	0.17
30	10	0.20
40	13	0.18
18	7	0.14
10	5	0.08
15	5	0.31
20	7	0.21

TABLE 2
Hypothetical replications of the aspirin/heart attack study

	Replication #1 Heart attack		Replication #2 Heart attack	
	Yes	No	Yes	No
Aspirin	11	1156	20	2314
Placebo	19	1090	48	2170
Chi-square	2.596, $p = 0.11$		13.206, $p = 0.0003$	

very *different* proportions, and in fact the relative risk from the second study is not even contained in a 95% confidence interval for relative risk from the original study. The *magnitude* of the effect has been much more closely matched by the "nonsignificant" replication.

Fortunately, psychologists are beginning to notice that replication is not as straightforward as they were originally led to believe. A special issue of the *Journal of Social Behavior and Personality* was entirely devoted to the question of replication (Neuliep, 1990). In one of the articles, Rosenthal cautioned his colleagues: "Given the levels of statistical power at which we normally operate, we have no right to expect the proportion of significant results that we typically do expect, even if in nature there is a very real and very important effect" (Rosenthal, 1990b, page 16).

Jacob Cohen, in his insightful article titled "Things I Have Learned (So Far)," identified another misconception common among social scientists: "Despite widespread misconceptions to the contrary, the rejection of a given null hypothesis gives us no basis for estimating the probability that a replication of the research will again result in rejecting that null hypothesis" (Cohen, 1990, page 1307).

Cohen and Rosenthal both advocate the use of effect sizes as opposed to significance levels when defining the strength of an experimental effect. In general, effect sizes measure the amount by which the data deviate from the null hypothesis in terms of standardized units. For instance, the effect size for a two-sample t-test is usually defined to be the difference in the two means, divided by the standard deviation for the control group. This measure can be compared across studies without the dependence on sample size inherent in significance levels. (Of course there will still be variability in the sample effect sizes, decreasing as a function of sample size.) Comparison of effect sizes across studies is one of the major components of meta-analysis.

Similar arguments have recently been made in the medical literature. For example, Gardner and Altman (1986) stated that the use of p-values "to define two alternative outcomes—significant and not significant—is not helpful and encourages lazy thinking" (page 746). They advocated the use of confidence intervals instead.

As discussed in the next section, the arguments used to conclude that parapsychology has failed to demonstrate a replicable effect hinge on these misconceptions of replication and failure to examine power. A more appropriate analysis would compare the effect sizes for similar experiments across experimenters and across time to see if there have been consistent effects of the same magnitude. Rosenthal also advocates this view of replication:

> The traditional view of replication focuses on significance level as the relevant summary statistic of a study and evaluates the success of a replication in a dichotomous fashion. The newer, more useful view of replication focuses on effect size as the more important summary statistic of a study and evaluates the success of a replication not in a dichotomous but in a continuous fashion [Rosenthal, 1990b, page 28].

The dichotomous view of replication has been used throughout the history of parapsychology, by both parapsychologists and critics (Utts, 1988). For example, the National Academy of Sciences report critically evaluated "significant" experiments, but entirely ignored "nonsignificant" experiments.

In the next three sections, we will examine some of the results in parapsychology using the broader, more appropriate definition of replication. In doing so, we will show that the results are far more interesting than the critics would have us believe.

4. THE GANZFELD DEBATE IN PARAPSYCHOLOGY

An extensive debate took place in the mid-1980s between a parapsychologist and critic, questioning whether or not a particular body of parapsychological data had demonstrated psi abilities. The experiments in question were all conducted using the ganzfeld setting (described below). Several authors were invited to write commentaries on the debate. As a result, this data base has been more thoroughly analyzed by both critics and proponents than any other and provides a good source for studying replication in parapsychology.

The debate concluded with a detailed series of recommendations for further experiments, and left open the question of whether or not psi abilities had been demonstrated. A new series of experiments that followed the recommendations were conducted over the next few years. The results of the new experiments will be presented in Section 5.

4.1 Free-Response Experiments

Recent experiments in parapsychology tend to use more complex target material than the cards and dice used in the early investigations, partially to alleviate boredom on the part of the subjects and partially because they are thought to "more nearly resemble the conditions of spontaneous psi occurrences" (Burdick and Kelly, 1977, page 109). These experiments fall under the general heading of "free-response" experiments, because the subject is asked to give a verbal or written description of the target, rather than being forced to make a choice from a small discrete set of possibilities. Various types of target material have been used, including pictures, short segments of movies on video tapes, actual locations and small objects.

Despite the more complex target material, the statistical methods used to analyze these experiments are similar to those for forced-choice experiments. A typical experiment proceeds as follows. Before conducting any trials, a large pool of potential targets is assembled, usually in packets of four. Similarity of targets within a packet is kept to a minimum, for reasons made clear below. At the start of an experimental session, after the subject is sequestered in an isolated room, a target is selected at random from the pool. A sender is placed in another room with the target. The subject is asked to provide a verbal or written description of what he or she thinks is in the target, knowing only that it is a photograph, an object, etc.

After the subject's description has been recorded and secured against the potential for later alteration, a judge (who may or may not be the subject) is given a copy of the subject's description and the four possible targets that were in the packet with the correct target. A properly conducted experiment either uses video tapes or has two identical sets of target material and uses the duplicate set for this part of the process, to ensure that clues such as fingerprints don't give away the answer. Based on the subject's description, and of course on a blind basis, the judge is asked to either rank the four choices from most to least likely to have been the target, or to select the one from the four that seems to best match the subject's description. If ranks are used, the statistical analysis proceeds by summing the ranks over a series of trials and comparing the sum to what would be expected by chance. If the selection method is used, a "direct hit" occurs if the correct target is chosen, and the number of direct hits over a series of trials is compared to the number expected in a binomial experiment with $p = 0.25$.

Note that the subjects' *responses* cannot be considered to be "random" in any sense, so probability assessments are based on the random selection of the target and decoys. In a correctly designed experiment, the probability of a direct hit by chance is 0.25 on each trial, regardless of the response, and the trials are independent. These and other issues related to analyzing free-response experiments are discussed by Utts (1991).

4.2 The Psi Ganzfeld Experiments

The ganzfeld procedure is a particular kind of free-response experiment utilizing a perceptual isolation technique originally developed by Gestalt psychologists for other purposes. Evidence from spontaneous case studies and experimental work had led parapsychologists to a model proposing that psychic functioning may be masked by sensory input and by inattention to internal states (Honorton, 1977). The ganzfeld procedure was specifically designed to test whether or not reduction of external "noise" would enhance psi performance.

In these experiments, the subject is placed in a comfortable reclining chair in an acoustically shielded room. To create a mild form of sensory deprivation, the subject wears headphones through which white noise is played, and stares into a constant field of red light. This is achieved by taping halved translucent ping-pong balls over the eyes and then illuminating the room with red light. In the psi ganzfeld experiments, the subject speaks into a microphone and attempts to describe the target material being observed by the sender in a distant room.

At the 1982 Annual Meeting of the Parapsychological Association, a debate took place over the degree to which the results of the psi ganzfeld experiments constituted evidence of psi abilities. Psychologist and critic Ray Hyman and parapsychologist Charles Honorton each analyzed the results of all known psi ganzfeld experiments to date, and they reached strikingly different conclusions (Honorton, 1985b; Hyman, 1985b). The debate continued with the publication of their arguments in separate articles in the March 1985 issue of the *Journal of Parapsychology*. Finally, in the December 1986 issue of the *Journal of Parapsychology*, Hyman and Honorton (1986) wrote a joint article in which they highlighted their agreements and disagreements and outlined detailed criteria for future experiments. That same issue contained commentaries on the debate by 10 other authors.

The data base analyzed by Hyman and Honorton (1986) consisted of results taken from 34 reports written by a total of 47 authors. Honorton counted 42 separate experiments described in the reports, of which 28 reported enough information to determine the number of direct hits achieved. Twenty three of the studies (55%) were classified by Honorton as having achieved statistical significance at 0.05.

4.3 The Vote-Counting Debate

Vote-counting is the term commonly used for the technique of drawing inferences about an experimental effect by counting the number of significant versus nonsignificant studies of the effect. Hedges and Olkin (1985) give a detailed analysis of the inadequacy of this method, showing that it is more and more likely to make the wrong decision as the number of studies increases. While Hyman acknowledged that "vote-counting raises many problems" (Hyman, 1985b, page 8), he nonetheless spent half of his critique of the ganzfeld studies showing why Honorton's count of 55% was wrong.

Hyman's first complaint was that several of the studies contained multiple conditions, each of which should be considered as a separate study. Using this definition he counted 80 studies (thus further reducing the sample sizes of the individual studies), of which 25 (31%) were "successful." Honorton's response to this was to invite readers to examine the studies and decide for themselves if the varying conditions constituted separate experiments.

Hyman next postulated that there was selection bias, so that significant studies were more likely to be reported. He raised some important issues about how pilot studies may be terminated and not reported if they don't show significant results, or may at least be subject to optional stopping, allowing the experimenter to determine the number of trials. He also presented a chi-square analysis that "suggests a tendency to report studies with a small sample only if they have significant results" (Hyman, 1985b, page 14), but I have questioned his analysis elsewhere (Utts, 1986, page 397).

Honorton refuted Hyman's argument with four rejoinders (Honorton, 1985b, page 66). In addition to reinterpreting Hyman's chi-square analysis, Honorton pointed out that the Parapsychological Association has an official policy encouraging the publication of nonsignificant results in its journals and proceedings, that a large number of reported ganzfeld studies did not achieve statistical significance and that there would have to be 15 studies in the "file-drawer" for every one reported to cancel out the observed significant results.

The remainder of Hyman's vote-counting analysis consisted of showing that the effective error rate for each study was actually much higher than the nominal 5%. For example, each study could have been analyzed using the direct hit measure, the sum of ranks measure or one of two other measures used for free-response analyses. Hyman carried out a simulation study that showed the true error rate would be 0.22 if "significance" was defined by requiring at least one of these four measures to achieve the 0.05 level. He suggested several other ways in which multiple testing could occur and concluded that the effective error rate in each experiment was not the nominal 0.05, but rather was probably close to the 31% he had determined to be the actual success rate in his vote-count.

Honorton acknowledged that there was a multiple testing problem, but he had a two-fold response. First, he applied a Bonferroni correction and found that the number of significant studies (using his definition of a study) only dropped from 55% to 45%. Next, he proposed that a uniform index of success be applied to all studies. He used the number of direct hits, since it was by far the most commonly reported measure and was the measure used in the first published psi ganzfeld study. He then conducted a detailed analysis of the 28 studies reporting direct hits and found that 43% were significant at 0.05 on that measure alone. Further, he showed that significant effects were reported by six of the 10 independent investigators and thus were not due to just one or two investigators or laboratories. He also noted that success rates were very similar for reports published in refereed journals and those published in unrefereed monographs and abstracts.

While Hyman's arguments identified issues such as selective reporting and optional stopping that should be considered in any meta-analysis, the dependence of significance levels on sample size makes the vote-counting technique almost useless for assessing the magnitude of the effect. Consider, for example, the 24 studies where the direct hit measure was reported and the chance probability of a direct hit was 0.25, the most common type of study in the data base. (There were four direct hit studies with other chance probabilities and 14 that did not report direct hits.) Of the 24 studies, 13 (54%) were "nonsignificant" at $\alpha = 0.05$, one-tailed. But if the 367 trials in these "failed replications" are combined, there are 106 direct hits, $z = 1.66$, and $p = 0.0485$, one tailed. This is reminiscent of the dilemma of Professor B in Section 3.

Power is typically very low for these studies. The median sample size for the studies reporting direct hits was 28. If there is a real effect and it increases the success probability from the chance 0.25 to an actual 0.33 (a value whose rationale will be made clear below), the power for a study with 28 trials is only 0.181 (Utts, 1986). It should be no surprise that there is a "repeatability" problem in parapsychology.

4.4 Flaw Analysis and Future Recommendations

The second half of Hyman's paper consisted of a "Meta-Analysis of Flaws and Successful Outcomes" (1985b, page 30), designed to explore whether or not various measures of success were related to specific flaws in the experiments. While many critics have argued that the results in parapsychology can be explained by experimental flaws, Hyman's analysis was the first to attempt to quantify the relationship between flaws and significant results.

Hyman identified 12 potential flaws in the ganzfeld experiments, such as inadequate randomization, multiple tests used without adjusting the significance level (thus inflating the significance level from the nominal 5%) and failure to use a duplicate set of targets for the judging process (thus allowing possible clues such as fingerprints). Using cluster and factor analyses, the 12 binary flaw variables were combined into three new variables, which Hyman named General Security, Statistics and Controls.

Several analyses were then conducted. The one reported with the most detail is a factor analysis utilizing 17 variables for each of 36 studies. Four factors emerged from the analysis. From these, Hyman concluded that security had increased over the years, that the significance level tended to be inflated the most for the most complex studies and that both effect size and level of significance were correlated with the existence of flaws.

Following his factor analysis, Hyman picked the three flaws that seemed to be most highly correlated with success, which were inadequate attention to both randomization and documentation and the potential for ordinary communication between the sender and receiver. A regression equation was then computed using each of the three flaws as dummy variables, and the effect size for the experiment as the dependent variable. From this equation, Hyman concluded that a study without these three flaws would be predicted to have a hit rate of 27%. He concluded that this is "well within the statistical neighborhood of the 25% chance rate" (1985b, page 37), and thus "the ganzfeld psi data base, despite initial impressions, is inadequate either to support the contention of a repeatable study or to demonstrate the reality of psi" (page 38).

Honorton discounted both Hyman's flaw classification and his analysis. He did not deny that flaws existed, but he objected that Hyman's analysis was faulty and impossible to interpret. Honorton asked psychometrician David Saunders to write an Appendix to his article, evaluating Hyman's analysis. Saunders first criticized Hyman's use of a factor analysis with 17 variables (many of which were dichotomous) and only 36 cases and concluded that "the entire analysis is meaningless" (Saunders, 1985, page 87). He then noted that Hyman's choice of the three flaws to include in his regression analysis constituted a clear case of multiple analysis, since there were 84 possible sets of three that could have been selected (out of nine potential flaws), and Hyman chose the set most highly correlated with effect size. Again, Saunders concluded that "any interpretation drawn from [the regression analysis] must be regarded as meaningless" (1985, page 88).

Hyman's results were also contradicted by Harris and Rosenthal (1988b) in an analysis requested by Hyman in his capacity as Chair of the National Academy of Sciences' Subcommittee on Parapsychology. Using Hyman's flaw classifications and a multivariate analysis, Harris and Rosenthal concluded that "Our analysis of the effects of flaws on study outcome lends no support to the hypothesis that ganzfeld research results are a significant function of the set of flaw variables" (1988b, page 3).

Hyman and Honorton were in the process of preparing papers for a second round of debate when they were invited to lunch together at the 1986 Meeting of the Parapsychological Association. They discovered that they were in general agreement on several major issues, and they decided to coauthor a "Joint Communiqué" (Hyman and Honorton, 1986). It is clear from their paper that they both thought it was more important to set the stage for future experimentation than to continue the technical arguments over the current data base. In the abstract to their paper, they wrote:

> We agree that there is an overall significant effect in this data base that cannot reasonably be explained by selective reporting or multiple analysis. We continue to differ over the degree to which the effect constitutes evidence for psi, but we agree that the final verdict awaits the outcome of future experiments conducted by a broader range of investigators and according to more stringent standards [page 351].

The paper then outlined what these standards should be. They included controls against any kind of sensory leakage, thorough testing and documentation of randomization methods used, better reporting of judging and feedback protocols, control for multiple analyses and advance specification of number of trials and type of experiment. Indeed, any area of research could benefit from such a careful list of procedural recommendations.

4.5 Rosenthal's Meta-Analysis

The same issue of the *Journal of Parapsychology* in which the Joint Communiqué appeared also carried commentaries on the debate by 10 separate authors. In his commentary, psychologist Robert Rosenthal, one of the pioneers of meta-analysis in psychology, summarized the aspects of Hyman's and Honorton's work that would typically be included in a meta-analysis (Rosenthal, 1986). It is worth reviewing Rosenthal's results so that they can be used as a basis of comparison for the more recent psi ganzfeld studies reported in Section 5.

Rosenthal, like Hyman and Honorton, focused only on the 28 studies for which direct hits were known. He chose to use an effect size measure called Cohen's h, which is the difference between the arcsin transformed proportions of direct hits that were observed and expected:

$$h = 2(\arcsin\sqrt{\hat{p}} - \arcsin\sqrt{p}).$$

One advantage of this measure over the difference in raw proportions is that it can be used to compare experiments with different chance hit rates.

If the observed and expected numbers of hits were identical, the effect size would be zero. Of the 28 studies, 23 (82%) had effect sizes greater than zero, with a median effect size of 0.32 and a mean of 0.28. These correspond to direct hit rates of 0.40 and 0.38 respectively, when 0.25 is expected by chance. A 95% confidence interval for the true effect size is from 0.11 to 0.45, corresponding to direct hit rates of from 0.30 to 0.46 when chance is 0.25.

A common technique in meta-analysis is to calculate a "combined z," found by summing the individual z scores and dividing by the square root of the number of studies. The result should have a standard normal distribution if each z score has a standard normal distribution. For the ganzfeld studies, Rosenthal reported a combined z of 6.60 with a p-value of 3.37×10^{-11}. He also reiterated Honorton's file-drawer assessment by calculating that there would have to be 423 studies unreported to negate the significant effect in the 28 direct hit studies.

Finally, Rosenthal acknowledged that, because of the flaws in the data base and the potential for at least a small file-drawer effect, the true average effect size was probably closer to 0.18 than 0.28. He concluded, "Thus, when the accuracy rate expected under the null is 1/4, we might estimate the obtained accuracy rate to be about 1/3" (1986, page 333). This is the value used for the earlier power calculation.

It is worth mentioning that Rosenthal was commissioned by the National Academy of Sciences to prepare a background paper to accompany its 1988 report on parapsychology. That paper (Harris and Rosenthal, 1988a) contained much of the same analysis as his commentary summarized above. Ironically, the discussion of the ganzfeld work in the National Academy Report focused on Hyman's 1985 analysis, but never mentioned the work it had commissioned Rosenthal to perform, which contradicted the final conclusion in the report.

5. A META-ANALYSIS OF RECENT GANZFELD EXPERIMENTS

After the initial exchange with Hyman at the 1982 Parapsychological Association Meeting, Honorton and his colleagues developed an automated ganzfeld experiment that was designed to eliminate the methodological flaws identified by Hyman. The execution and reporting of the experiments followed the detailed guidelines agreed upon by Hyman and Honorton.

Using this "autoganzfeld" experiment, 11 experimental series were conducted by eight experimenters between February 1983 and September 1989, when the equipment had to be dismantled due to lack of funding. In this section, the results of these experiments are summarized and compared to the earlier ganzfeld studies. Much of the information is derived from Honorton et al. (1990).

5.1 The Automated Ganzfeld Procedure

Like earlier ganzfeld studies, the "autoganzfeld" experiments require four participants. The first is the Receiver (R), who attempts to identify the target material being observed by the Sender (S). The Experimenter (E) prepares R for the task, elicits the response from R and supervises R's judging of the response against the four potential targets. (Judging is double blind; E does not know which is the correct target.) The fourth participant is the lab assistant (LA) whose only task is to instruct the computer to randomly select the target. No one involved in the experiment knows the identity of the target.

Both R and S are sequestered in sound-isolated, electrically shielded rooms. R is prepared as in earlier ganzfeld studies, with white noise and a field of red light. In a nonadjacent room, S watches the target material on a television and can hear R's target description ("mentation") as it is being given. The mentation is also tape recorded.

The judging process takes place immediately after the 30-minute sending period. On a TV monitor in the isolated room, R views the four choices from the target pack that contains the actual target. R is asked to rate each one according to how closely it matches the ganzfeld mentation. The ratings are converted to ranks and, if the correct target is ranked first, a direct hit is scored. The entire process is automatically recorded by the computer. The computer then displays the correct choice to R as feedback.

There were 160 preselected targets, used with replacement, in 10 of the 11 series. They were arranged in packets of four, and the decoys for a given target were always the remaining three in the same set. Thus, even if a particular target in a set were consistently favored by Rs, the probability of a direct hit under the null hypothesis would remain at 1/4. Popular targets should be no more likely to be selected by the computer's random number generator than any of the others in the set. The selection of the target by the computer is the only source of randomness in these experiments. This is an important point, and one that is often misunderstood. (See Utts, 1991, for elucidation.)

Eighty of the targets were "dynamic," consisting of scenes from movies, documentaries and cartoons; 80 were "static," consisting of photographs, art prints and advertisements. The four targets within each set were all of the same type. Earlier studies indicated that dynamic targets were more likely to produce successful results, and one of the goals of the new experiments was to test that theory.

The randomization procedure used to select the target and the order of presentation for judging was thoroughly tested before and during the experiments. A detailed description is given by Honorton et al. (1990, pages 118–120).

Three of the 11 series were pilot series, five were formal series with novice receivers, and three were formal series with experienced receivers. The last series with experienced receivers was the only one that did not use the 160 targets. Instead, it used only one set of four dynamic targets in which one target had previously received several first place ranks and one had never received a first place rank. The receivers, none of whom had had prior exposure to that target pack, were not aware that only one target pack was being used. They each contributed one session only to the series. This will be called the "special series" in what follows.

Except for two of the pilot series, numbers of trials were planned in advance for each series. Unfortunately, three of the formal series were not yet completed when the funding ran out, including the special series, and one pilot study with advance planning was terminated early when the experimenter relocated. There were no unreported trials during the 6-year period under review, so there was no "file drawer."

Overall, there were 183 Rs who contributed only one trial and 58 who contributed more than one, for a total of 241 participants and 355 trials. Only 23 Rs had previously participated in ganzfeld experiments, and 194 Rs (81%) had never participated in any parapsychological research.

5.2 Results

While acknowledging that no probabilistic conclusions can be drawn from qualitative data, Honorton et al. (1990) included several examples of session excerpts that Rs identified as providing the basis for their target rating. To give a flavor for the dream-like quality of the mentation and the amount of information that can be lost by only assigning a rank, the first example is reproduced here. The target was a painting by Salvador Dali called "Christ Crucified." The correct target received a first place rank. The part of the mentation R used to make this assessment read:

> ... I think of guides, like spirit guides, leading me and I come into a court with a king. It's quiet.... It's like heaven. The king is something like Jesus. Woman. Now I'm just sort of summersaulting through heaven.... Brooding.... Aztecs, the Sun God.... High priest.... Fear.... Graves. Woman. Prayer.... Funeral.... Dark. Death.... Souls.... Ten Commandments. Moses.... [Honorton et al., 1990].

Over all 11 series, there were 122 direct hits in the 355 trials, for a hit rate of 34.4% (exact binomial p-value = 0.00005) when 25% were expected by chance. Cohen's h is 0.20, and a 95% confidence interval for the overall hit rate is from 0.30 to 0.39. This calculation assumes, of course, that the probability of a direct hit is constant and independent across trials, an assumption that may be questionable except under the null hypothesis of no psi abilities.

Honorton et al. (1990) also calculated effect sizes for each of the 11 series and each of the eight experimenters. All but one of the series (the first novice series) had positive effect sizes, as did all of the experimenters.

The special series with experienced Rs had an exceptionally high effect size with $h = 0.81$, corresponding to 16 direct hits out of 25 trials (64%), but the remaining series and the experimenters had relatively homogeneous effect sizes given the amount of variability expected by chance. If the special series is removed, the overall hit rate is 32.1%, $h = 0.16$. Thus, the positive effects are not due to just one series or one experimenter.

Of the 218 trials contributed by novices, 71 were direct hits (32.5%, $h = 0.17$), compared with 51 hits in the 137 trials by those with prior ganzfeld experience (37%, $h = 0.26$). The hit rates and effect sizes were 31% ($h = 0.14$) for the combined pilot series, 32.5% ($h = 0.17$) for the combined formal novice series, and 41.5% ($h = 0.35$) for the combined experienced series. The last figure drops to 31.6% if the outlier series is removed. Finally, without the outlier series the hit rate for the combined series where all of the planned trials were completed was 31.2% ($h = 0.14$), while it was 35% ($h = 0.22$) for the combined series that were terminated early. Thus, optional stopping cannot account for the positive effect.

There were two interesting comparisons that had been suggested by earlier work and were preplanned in these experiments. The first was to compare results for trials with dynamic targets with those for static targets. In the 190 dynamic target sessions there were 77 direct hits (40%, $h = 0.32$) and for the static targets there were 45 hits in 165 trials (27%, $h = 0.05$), thus indicating that dynamic targets produced far more successful results.

The second comparison of interest was whether or not the sender was a friend of the receiver. This was a choice the receiver could make. If he or she did not bring a friend, a lab member acted as sender. There were 211 trials with friends as senders (some of whom were also lab staff), resulting in 76 direct hits (36%, $h = 0.24$). Four trials used no sender. The remaining 140 trials used nonfriend lab staff as senders and resulted in 46 direct hits (33%, $h = 0.18$). Thus, trials with friends as senders were slightly more successful than those without.

Consonant with the definition of replication based on consistent effect sizes, it is informative to compare the autoganzfeld experiments with the direct hit studies in the previous data base. The overall success rates are extremely similar. The overall direct hit rate was 34.4% for the autoganzfeld studies and was 38% for the comparable direct hit studies in the earlier meta-analysis. Rosenthal's (1986) adjustment for flaws had placed a more conservative estimate at 33%, very close to the observed 34.4% in the new studies.

One limitation of this work is that the autoganzfeld studies, while conducted by eight experimenters, all used the same equipment in the same laboratory. Unfortunately, the level of funding available in parapsychology and the cost in time and equipment to conduct proper experiments make it difficult to amass large amounts of data across laboratories. Another autoganzfeld laboratory is currently being constructed at the University of Edinburgh in Scotland, so interlaboratory comparisons may be possible in the near future.

Based on the effect size observed to date, large samples are needed to achieve reasonable power. If there is a constant effect across all trials, resulting in 33% direct hits when 25% are expected by chance, to achieve a one-tailed significance level of 0.05 with 95% probability would require 345 sessions.

We end this section by returning to the aspirin and heart attack example in Section 3 and expanding a comparison noted by Atkinson, Atkinson, Smith and Bem (1990, page 237). Computing the equivalent of Cohen's h for comparing observed heart attack rates in the aspirin and placebo groups results in $h = 0.068$. Thus, the effect size observed in the ganzfeld data base is triple the much publicized effect of aspirin on heart attacks.

6. OTHER META-ANALYSES IN PARAPSYCHOLOGY

Four additional meta-analyses have been conducted in various areas of parapsychology since the original ganzfeld meta-analyses were reported. Three of the four analyses focused on evidence of psi abilities, while the fourth examined the relationship between extroversion and psychic functioning. In this section, each of the four analyses will be briefly summarized.

There are only a handful of English-language journals and proceedings in parapsychology, so retrieval of the relevant studies in each of the four cases was simple to accomplish by searching those sources in detail and by searching other bibliographic data bases for keywords.

Each analysis included an overall summary, an analysis of the quality of the studies versus the size of the effect and a "file-drawer" analysis to determine the possible number of unreported studies. Three of the four also contained comparisons across various conditions.

6.1 Forced-Choice Precognition Experiments

Honorton and Ferrari (1989) analyzed forced-choice experiments conducted from 1935 to 1987, in which the target material was randomly selected *after* the subject had attempted to predict what it would be. The time delay in selecting the target ranged from under a second to one year. Target material included items as diverse as ESP cards and automated random number generators. Two investigators, S. G. Soal and Walter J. Levy, were not included because some of their work has been suspected to be fraudulent.

Overall Results. There were 309 studies reported by 62 senior authors, including more than 50,000 subjects and nearly two million individual trials. Honorton and Ferrari used $z/\sqrt{n}$ as the measure of effect size (ES) for each study, where n was the number of Bernoulli trials in the study. They reported a mean ES of 0.020, and a mean z-score of 0.65 over all studies. They also reported a combined z of 11.41, $p = 6.3 \times 10^{-25}$. Some 30% (92) of the studies were statistically significant at $\alpha = 0.05$. The mean ES per investigator was 0.033, and the significant results were not due to just a few investigators.

Quality. Eight dichotomous quality measures were assigned to each study, resulting in possible scores from zero for the lowest quality, to eight for the highest. They included features such as adequate randomization, preplanned analysis and automated recording of the results. The correlation between study quality and effect size was 0.081, indicating a slight tendency for higher quality studies to be more successful, contrary to claims by critics that the opposite would be true. There was a clear relationship between quality and year of publication, presumably because over the years experimenters in parapsychology have responded to suggestions from critics for improving their methodology.

File Drawer. Following Rosenthal (1984), the authors calculated the "fail-safe N" indicating the number of unreported studies that would have to be sitting in file drawers in order to negate the significant effect. They found $N = 14{,}268$, or a ratio of 46 unreported studies for each one reported. They also followed a suggestion by Dawes, Landman and Williams (1984) and computed the mean z for all studies with $z > 1.65$. If such studies were a random sample from the upper 5% tail of a $N(0,1)$ distribution, the mean z would be 2.06. In this case it was 3.61. They concluded that selective reporting could not explain these results.

Comparisons. Four variables were identified that appeared to have a systematic relationship to study outcome. The first was that the 25 studies using subjects selected on the basis of good past performance were more successful than the 223 using unselected subjects, with mean effect sizes of 0.051 and 0.008, respectively. Second, the 97 studies testing subjects individually were more successful than the 105 studies that used group testing; mean effect sizes were 0.021 and 0.004, respectively. Timing of feedback was the third moderating variable, but information was only available for 104 studies. The 15 studies that never told the subjects what the targets were had a mean effect size of −0.001. Feedback after each trial produced the best results, the mean ES for the 47 studies was 0.035. Feedback after each set of trials resulted in mean ES of 0.023 (21 studies), while delayed feedback (also 21 studies) yielded a mean ES of only 0.009. There is a clear ordering; as the gap between time of feedback and time of the actual guesses decreased, effect sizes increased.

The fourth variable was the time interval between the subject's guess and the actual target selection, available for 144 studies. The best results were for the 31 studies that generated targets less than a second after the guess (mean $ES = 0.045$), while the worst were for the seven studies that delayed target selection by at least a month (mean $ES = 0.001$). The mean effect sizes showed a clear trend, decreasing in order as the time interval increased from minutes to hours to days to weeks to months.

6.2 Attempts to Influence Random Physical Systems

Radin and Nelson (1989) examined studies designed to test the hypothesis that "The statistical output of an electronic RNG [random number generator] is correlated with observer intention in accordance with prespecified instructions" (page 1502). These experiments typically involve RNGs based on radioactive decay, electronic noise or pseudorandom number sequences seeded with true random sources. Usually the subject is instructed to try to influence the results of a string of binary trials by mental intention alone. A typical protocol would ask a subject to press a button (thus starting the collection of a fixed-length sequence of bits), and then try to influence the random source to produce more zeroes or more ones. A run might consist of three successive button presses, one each in which the desired result was more zeroes or more ones, and one as a control with no conscious intention. A z score would then be computed for each button press.

The 832 studies in the analysis were conducted from 1959 to 1987 and included 235 "control" studies, in which the output of the RNGs were recorded but there was no conscious intention involved. These were usually conducted before and during the experimental series, as tests of the RNGs.

Results. The effect size measure used was again $z/\sqrt{n}$, where z was positive if more bits of the specified type were achieved. The mean effect size for control studies was not significantly different from zero (-1.0×10^{-5}). The mean effect size for the experimental studies was also very small, 3.2×10^{-4}, but it was significantly higher than the mean ES for the control studies ($z = 4.1$).

Quality. Sixteen quality measures were defined and assigned to each study, under the four general categories of procedures, statistics, data and the RNG device. A score of 16 reflected the highest quality. The authors regressed mean effect size on mean quality for each investigator and found a slope of 2.5×10^{-5} with standard error of 3.2×10^{-5}, indicating little relationship between quality and outcome. They also calculated a weighted mean effect size, using quality scores as weights, and found that it was very similar to the unweighted mean ES. They concluded that "differences in methodological quality are not significant predictors of effect size" (page 1507).

File Drawer. Radin and Nelson used several methods for estimating the number of unreported studies (pages 1508–1510). Their estimates ranged from 200 to 1000 based on models assuming that all significant studies were reported. They calculated the fail-safe N to be 54,000.

6.3 Attempts to Influence Dice

Radin and Ferrari (1991) examined 148 studies, published from 1935 to 1987, designed to test whether or not consciousness can influence the results of tossing dice. They also found 31 "control" studies in which no conscious intention was involved.

Results. The effect size measure used was $z/\sqrt{n}$, where z was based on the number of throws in which the die landed with the desired face (or faces) up, in n throws. The weighted mean ES for the experimental studies was 0.0122 with a standard error of 0.00062; for the control studies the mean and standard error were 0.00093 and 0.00255, respectively. Weights for each study were determined by quality, giving more weight to high-quality studies. Combined z scores for the experimental and control studies were reported by Radin and Ferrari to be 18.2 and 0.18, respectively.

Quality. Eleven dichotomous quality measures were assigned, ranging from automated recording to whether or not control studies were interspersed with the experimental studies. The final quality score for each study combined these with information on method of tossing the dice, and with source of subject (defined below). A regression of quality score versus effect size resulted in a slope of -0.002, with a standard error of 0.0011. However, when effect sizes were weighted by sample size, there was a significant relationship between quality and effect size, leading Radin and Ferrari to conclude that higher-quality studies produced lower weighted effect sizes.

File Drawer. Radin and Ferrari calculated Rosenthal's fail-safe N for this analysis to be 17,974. Using the assumption that all significant studies were reported, they estimated the number of unreported studies to be 1152. As a final assessment, they compared studies published before and after 1975, when the *Journal of Parapsychology* adopted an official policy of publishing nonsignificant results. They concluded, based on that analysis, that more nonsignificant studies were published after 1975, and thus "We must consider the overall (1935–1987) data base as suspect with respect to the filedrawer problem."

Comparisons. Radin and Ferrari noted that there was bias in both the experimental and control studies across die face. Six was the face most likely to come up, consistent with the observation that it has the least mass. Therefore, they examined results for the subset of 69 studies in which targets were evenly balanced among the six faces. They still found a significant effect, with mean and standard error for effect size of 8.6×10^{-3} and 1.1×10^{-3}, respectively. The combined z was 7.617 for these studies.

They also compared effect sizes across types of subjects used in the studies, categorizing them as unselected, experimenter and other subjects, experimenter as sole subject, and specially selected subjects. Like Honorton and Ferrari (1989), they found the highest mean ES for studies with selected subjects; it was approximately 0.02, more than twice that for unselected subjects.

6.4 Extroversion and ESP Performance

Honorton, Ferrari and Bem (1991) conducted a meta-analysis to examine the relationship between scores on tests of extroversion and scores on psi-related tasks. They found 60 studies by 17 investigators, conducted from 1945 to 1983.

Results. The effect size measure used for this analysis was the correlation between each subject's extroversion score and ESP score. A variety of measures had been used for both scores across studies, so various correlation coefficients were used. Nonetheless, a stem and leaf diagram of the correlations showed an approximate bell shape with mean and standard deviation of 0.19 and 0.26, respectively, and with an additional outlier at $r = 0.91$. Honorton et al. reported that when weighted by degrees of freedom, the weighted mean r was 0.14, with a 95% confidence interval covering 0.10 to 0.19.

Forced-Choice versus Free-Response Results. Because forced-choice and free-response tests differ qualitatively, Honorton et al. chose to examine their relationship to extroversion separately. They found that for free-response studies there was a significant correlation between extroversion and ESP scores, with mean $r = 0.20$ and $z = 4.46$. Further, this effect was homogeneous across both investigators and extroversion scales.

For forced-choice studies, there was a significant correlation between ESP and extroversion, but only for those studies that reported the ESP results to the subjects *before* measuring extroversion. Honorton et al. speculated that the relationship was an artifact, in which extroversion scores were temporarily inflated as a result of positive feedback on ESP performance.

Confirmation with New Data Following the extroversion/ESP meta-analysis, Honorton et al. attempted to confirm the relationship using the autoganzfeld data base. Extroversion scores based on the Myers–Briggs Type Indicator were available for 221 of the 241 subjects who had participated in autoganzfeld studies.

The correlation between extroversion scores and ganzfeld rating scores was $r = 0.18$, with a 95% confidence interval from 0.05 to 0.30. This is consistent with the mean correlation of $r = 0.20$ for free-response experiments, determined from the meta-analysis. These correlations indicate that extroverted subjects can produce higher scores in free-response ESP tests.

7. CONCLUSIONS

Parapsychologists often make a distinction between "proof-oriented research" and "process-oriented research." The former is typically conducted to test the hypothesis that psi abilities exist, while the latter is designed to answer questions about how psychic functioning works. Proof-oriented research has dominated the literature in parapsychology. Unfortunately, many of the studies used small samples and would thus be nonsignificant even if a moderate-sized effect exists.

The recent focus on meta-analysis in parapsychology has revealed that there are small but consistently nonzero effects across studies, experimenters and laboratories. The sizes of the effects in forced-choice studies appear to be comparable to those reported in some medical studies that had been heralded as breakthroughs. (See Section 5; also Honorton and Ferrari, 1989, page 301.) Free-response studies show effect sizes of far greater magnitude.

A promising direction for future process-oriented research is to examine the causes of individual differences in psychic functioning. The ESP/extroversion meta-analysis is a step in that direction.

In keeping with the idea of individual differences, Bayes and empirical Bayes methods would appear to make more sense than the classical inference methods commonly used, since they would allow individual abilities and beliefs to be modeled. Jeffreys (1990) reported a Bayesian analysis of some of the RNG experiments and showed that conclusions were closely tied to prior beliefs even though hundreds of thousands of trials were available.

It may be that the nonzero effects observed in the meta-analyses can be explained by something other than ESP, such as shortcomings in our understanding of randomness and independence. Nonetheless, there is an anomaly that needs an explanation. As I have argued elsewhere (Utts, 1987), research in parapsychology should receive more support from the scientific community. If ESP does not exist, there is little to be lost by erring in the direction of further research, which may in fact uncover other anomalies. If ESP does exist, there is much to be lost by not doing process-oriented research and much to be gained by discovering how to enhance and apply these abilities to important world problems.

ACKNOWLEDGMENTS

I would like to thank Deborah Delanoy, Charles Honorton, Wesley Johnson, Scott Plous and an anonymous reviewer for their helpful comments on an earlier draft of this paper, and Robert Rosenthal and Charles Honorton for discussions that helped clarify details.

REFERENCES

Atkinson, R. L., Atkinson, R. C., Smith, E. E. and Bem, D. J. (1990). *Introduction to Psychology*, 10th ed. Harcourt Brace Jovanovich, San Diego.

Beloff, J. (1985). Research strategies for dealing with unstable phenomena. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 1–21. Parapsychology Foundation, New York.

Blackmore, S. J. (1985). Unrepeatability: Parapsychology's only finding. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 183–206. Parapsychology Foundation, New York.

Burdick, D. S. and Kelly, E. F. (1977). Statistical methods in parapsychological research. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 81–130. Van Nostrand Reinhold, New York.

Camp, B. H. (1937). (Statement in Notes Section.) *Journal of Parapsychology* 1 305.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist* 45 1304–1312.

Coover, J. E. (1917). *Experiments in Psychical Research at Leland Stanford Junior University*. Stanford Univ.

Dawes, R. M., Landman, J. and Williams, J. (1984). Reply to Kurosawa. *American Psychologist* 39 74–75.

Diaconis, P. (1978). Statistical problems in ESP research. *Science* 201 131–136.

Dommeyer, F. C. (1975). Psychical research at Stanford University. *Journal of Parapsychology* 39 173–205.

Druckman, D. and Swets, J. A., eds. (1988) *Enhancing Human Performance: Issues, Theories, and Techniques*. National Academy Press, Washington, D.C.

Edgeworth, F. Y. (1885). The calculus of probabilities applied to psychical research. In *Proceedings of the Society for Psychical Research* 3 190–199.

Edgeworth, F. Y. (1886). The calculus of probabilities applied to psychical research. II. In *Proceedings of the Society for Psychical Research* 4 189–208.

Feller, W. K. (1940). Statistical aspects of ESP. *Journal of Parapsychology* 4 271–297.

Feller, W. K. (1968). *An Introduction to Probability Theory and Its Applications* 1, 3rd ed. Wiley, New York.

Fisher, R. A. (1924). A method of scoring coincidences in tests with playing cards. In *Proceedings of the Society for Psychical Research* 34 181–185.

Fisher, R. A. (1929). The statistical method in psychical research. In *Proceedings of the Society for Psychical Research* 39 189–192.

Gallup, G. H., Jr., and Newport, F. (1991). Belief in paranormal phenomena among adult Americans. *Skeptical Inquirer* 15 137–146.

Gardner, M. J. and Altman, D. G. (1986). Confidence intervals rather than p-values: Estimation rather than hypothesis testing. *British Medical Journal* 292 746–750.

GILMORE, J. B. (1989). Randomness and the search for psi. *Journal of Parapsychology* 53 309–340.

GILMORE, J. B. (1990). Anomalous significance in pararandom and psi-free domains. *Journal of Parapsychology* 54 53–58.

GREELEY, A. (1987). Mysticism goes mainstream. *American Health* 7 47–49.

GREENHOUSE, J. B. and GREENHOUSE, S. W. (1988). An aspirin a day...? *Chance* 1 24–31.

GREENWOOD, J. A. and STUART, C. E. (1940). A review of Dr. Feller's critique. *Journal of Parapsychology* 4 299–319.

HACKING, I. (1988). Telepathy: Origins of randomization in experimental design. *Isis* 79 427–451.

HANSEL, C. E. M. (1980). *ESP and Parapsychology: A Critical Re-evaluation*. Prometheus Books, Buffalo, N.Y.

HARRIS, M. J. and ROSENTHAL, R. (1988a). *Interpersonal Expectancy Effects and Human Performance Research*. National Academy Press, Washington, D.C.

HARRIS, M. J. and ROSENTHAL, R. (1988b). *Postscript to Interpersonal Expectancy Effects and Human Performance Research*. National Academy Press, Washington, D.C.

HEDGES, L. V. and OLKIN, I. (1985). *Statistical Methods for Meta-Analysis*. Academic, Orlando, Fla.

HONORTON, C. (1977). Psi and internal attention states. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 435–472. Van Nostrand Reinhold, New York.

HONORTON, C. (1985a). How to evaluate and improve the replicability of parapsychological effects. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 238–255. Parapsychology Foundation, New York.

HONORTON, C. (1985b). Meta-analysis of psi ganzfeld research: A response to Hyman. *Journal of Parapsychology* 49 51–91.

HONORTON, C., BERGER, R. E., VARVOGLIS, M. P., QUANT, M., DERR, P., SCHECHTER, E. I. and FERRARI, D. C. (1990). Psi communication in the ganzfeld: Experiments with an automated testing system and a comparison with a meta-analysis of earlier studies. *Journal of Parapsychology* 54 99–139.

HONORTON, C. and FERRARI, D. C. (1989). "Future telling": A meta-analysis of forced-choice precognition experiments, 1935–1987. *Journal of Parapsychology* 53 281–308.

HONORTON, C., FERRARI, D. C. and BEM, D. J. (1991). Extraversion and ESP performance: A meta-analysis and a new confirmation. *Research in Parapsychology 1990*. The Scarecrow Press, Metuchen, N.J. To appear.

HYMAN, R. (1985a). A critical overview of parapsychology. In *A Skeptic's Handbook of Parapsychology* (P. Kurtz, ed.) 1–96. Prometheus Books, Buffalo, N.Y.

HYMAN, R. (1985b). The ganzfeld psi experiment: A critical appraisal. *Journal of Parapsychology* 49 3–49.

HYMAN, R. and HONORTON, C. (1986). Joint communiqué: The psi ganzfeld controversy. *Journal of Parapsychology* 50 351–364.

IVERSEN, G. R., LONGCOR, W. H., MOSTELLER, F., GILBERT, J. P. and YOUTZ, C. (1971). Bias and runs in dice throwing and recording: A few million throws. *Psychometrika* 36 1–19.

JEFFREYS, W. H. (1990). Bayesian analysis of random event generator data. *Journal of Scientific Exploration* 4 153–169.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* 44 187–192.

MAUSKOPF, S. H. and MCVAUGH, M. (1979). *The Elusive Science: Origins of Experimental Psychical Research*. Johns Hopkins Univ. Press.

MCVAUGH, M. R. and MAUSKOPF, S. H. (1976). J. B. Rhine's *Extrasensory Perception* and its background in psychical research. *Isis* 67 161–189.

NEULIEP, J. W., ed. (1990). Handbook of replication research in the behavioral and social sciences. *Journal of Social Behavior and Personality* 5 (4) 1–510.

OFFICE OF TECHNOLOGY ASSESSMENT (1989). Report of a workshop on experimental parapsychology. *Journal of the American Society for Psychical Research* 83 317–339.

PALMER, J. (1989). A reply to Gilmore. *Journal of Parapsychology* 53 341–344.

PALMER, J. (1990). Reply to Gilmore: Round two. *Journal of Parapsychology* 54 59–61.

PALMER, J. A., HONORTON, C. and UTTS, J. (1989). Reply to the National Research Council study on parapsychology. *Journal of the American Society for Psychical Research* 83 31–49.

RADIN, D. I. and FERRARI, D. C. (1991). Effects of consciousness on the fall of dice: A meta-analysis. *Journal of Scientific Exploration* 5 61–83.

RADIN, D. I. and NELSON, R. D. (1989). Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics* 19 1499–1514.

RAO, K. R. (1985). Replication in conventional and controversial sciences. In *The Repeatability Problem in Parapsychology* (B. Shapin and L. Coly, eds.) 22–41. Parapsychology Foundation, New York.

RHINE, J. B. (1934). *Extrasensory Perception*. Boston Society for Psychical Research, Boston. (Reprinted by Branden Press, 1964.)

RHINE, J. B. (1977). History of experimental studies. In *Handbook of Parapsychology* (B. B. Wolman, ed.) 25–47. Van Nostrand Reinhold, New York.

RICHET, C. (1884). La suggestion mentale et le calcul des probabilités. *Revue Philosophique* 18 608–674.

ROSENTHAL, R. (1984). *Meta-Analytic Procedures for Social Research*. Sage, Beverly Hills.

ROSENTHAL, R. (1986). Meta-analytic procedures and the nature of replication: The ganzfeld debate. *Journal of Parapsychology* 50 315–336.

ROSENTHAL, R. (1990a). How are we doing in soft psychology? *American Psychologist* 45 775–777.

ROSENTHAL, R. (1990b). Replication in behavioral research. *Journal of Social Behavior and Personality* 5 1–30.

SAUNDERS, D. R. (1985). On Hyman's factor analysis. *Journal of Parapsychology* 49 86–88.

SHAPIN, B. and COLY, L., eds. (1985). *The Repeatability Problem in Parapsychology*. Parapsychology Foundation, New York.

SPENCER-BROWN, G. (1957). *Probability and Scientific Inference*. Longmans Green, London and New York.

STUART, C. E. and GREENWOOD, J. A. (1937). A review of criticisms of the mathematical evaluation of ESP data. *Journal of Parapsychology* 1 295–304.

TVERSKY, A. and KAHNEMAN, D. (1982). Belief in the law of small numbers. In *Judgment Under Uncertainty: Heuristics and Biases* (D. Kahneman, P. Slovic and A. Tversky, eds.) 23–31. Cambridge Univ. Press.

UTTS, J. (1986). The ganzfeld debate: A statistician's perspective. *Journal of Parapsychology* 50 395–402.

UTTS, J. (1987). Psi, statistics, and society. *Behavioral and Brain Sciences* 10 615–616.

UTTS, J. (1988). Successful replication versus statistical significance. *Journal of Parapsychology* 52 305–320.

UTTS, J. (1989). Randomness and randomization tests: A reply to Gilmore. *Journal of Parapsychology* 53 345–351.

UTTS, J. (1991). Analyzing free-response data: A progress report. In *Psi Research Methodology: A Re-examination* (L. Coly, ed.). Parapsychology Foundation, New York. To appear.

WILKS, S. S. (1965a). Statistical aspects of expeirments in telepath. *N.Y. Statistician* 16 (6) 1–3.

WILKS, S. S. (1965b). Statistical aspects of experiments in telepathy. *N.Y. Statistician* 16 (7) 4–6.

Comment

M. J. Bayarri and James Berger

1. INTRODUCTION

There are many fascinating issues discussed in this paper. Several concern parapsychology itself and the interpretation of statistical methodology therein. We are not experts in parapsychology, and so have only one comment concerning such matters: In Section 3 we briefly discuss the need to switch from P-values to Bayes factors in discussing evidence concerning parapsychology.

A more general issue raised in the paper is that of replication. It is quite illuminating to consider the issue of replication from a Bayesian perspective, and this is done in Section 2 of our discussion.

2. REPLICATION

Many insightful observations concerning replication are given in the article, and these spurred us to determine if they could be quantified within Bayesian reasoning. Quantification requires clear delineation of the possible purposes of replication, and at least two are obvious. The first is simple reduction of random error, achieved by obtaining more observations from the replication. The second purpose is to search for possible bias in the original experiment. We use "bias" in a loose sense here, to refer to any of the huge number of ways in which the effects being measured by the experiment can differ from the actual effects of interest. Thus a clinical trial without a placebo can suffer a placebo "bias"; a survey can suffer a "bias" due to the sampling frame being unrepresentative of the actual population; and possible sources of bias in parapsychological experiments have been extensively discussed.

Replication to Reduce Random Error

If the sole goal of replication of an experiment is to reduce random error, matters are very straightforward. Reviewing the Bayesian way of studying this issue is, however, useful and will be done through the following simple example.

M. J. Bayarri is Titular Professor, Department of Statistics and Operations Research, University of Valencia, Avenida Dr. Moliner 50, 46100 Burjassot, Valencia, Spain. James Berger is the Richard M. Brumfield Distinguished Professor of Statistics,

EXAMPLE 1. Consider the example from Tversky and Kahnemann (1982), in which an experiment results in a standardized test statistic of $z_1 = 2.46$. (We will assume normality to keep computations trivial.) The question is: What is the highest value of z_2 in a second set of data that would be considered a failure to replicate? Two possible precise versions of this question are: Question 1: What is the probability of observing z_2 for which the null hypothesis would be rejected in the replicated experiment? Question 2: What value of z_2 would leave one's overall opinion about the null hypothesis unchanged?

Consider the simple case where $Z_1 \sim N(z_1 \mid \theta, 1)$ and (independently) $Z_2 \sim N(z_2 \mid \theta, 1)$, where θ is the mean and 1 is the standard deviation of the normal distribution. Note that we are considering the case in which no experimental bias is suspected and so the means for each experiment are assumed to be the same.

Suppose that it is desired to test $H_0: \theta \leq 0$ versus $H_1: \theta > 0$, and suppose that initial prior opinion about θ can be described by the noninformative prior $\pi(\theta) = 1$. We consider the one-sided testing problem with a constant prior in this section, because it is known that then the posterior probability of H_0, to be denoted by $P(H_0 \mid \text{data})$, equals the P-value, allowing us to avoid complications arising from differences between Bayesian and classical answers.

After observing $z_1 = 2.46$, the posterior distribution of θ is

$$\pi(\theta \mid z_1) = N(\theta \mid 2.46, 1).$$

Question 1 then has the answer (using predictive Bayesian reasoning)

$$P(\text{rejecting at level } \alpha \mid z_1) = \int_{c_\alpha}^{\infty} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z_2-\theta)^2} \pi(\theta \mid z_1)\, d\theta\, dz_2 = 1 - \Phi\left(\frac{c_\alpha - 2.46}{\sqrt{2}}\right),$$

where Φ is the standard normal cdf and c_α is the (one-sided) critical value corresponding to the level, α, of the test. For instance, if $\alpha = 0.05$, then this probability equals 0.7178, demonstrating that there is a quite substantial probability that the second experiment will fail to reject. If α is chosen to be the observed significance level from the first expersecond experiment will reject is just 1/2. This is nothing but a statement of the well-known martingale property of Bayesianism, that what you "expect" to see in the future is just what you know today. In a sense, therefore, question 1 is exposed as being uninteresting.

Question 2 more properly focuses on the fact that the stated goal of replication here is simply to reduce uncertainty in stated conclusions. The answer to the question follows immediately from noting that the posterior from the combined data (z_1, z_2) is

$$\pi(\theta \mid z_1, z_2) = N(\theta \mid (z_1 + z_2)/2, 1/\sqrt{2}),$$

so that

$$P(H_0 \mid \text{data}) = \Phi(-(z_1 + z_2)/\sqrt{2}).$$

Setting this equal to $P(H_0 \mid z_1)$ and solving for z_2 yields $z_2 = (\sqrt{2} - 1)z_1 = 1.02$. Any value of z_2 greater than this will increase the total evidence against H_0, while any value smaller than 1.02 will decrease the evidence.

Replication to Detect Bias

The aspirin example dramatically raises the issue of bias detection as a motive for replication. Professor Utts observes that replication 1 gives results that are fully compatible with those of the original study, which could be interpreted as suggesting that there is no bias in the original study, while replication 2 would raise serious concerns of bias. We became very interested in the implicit suggestion that replication 2 would thus lead to less overall evidence against the null hypothesis than would replication 1, even though in isolation replication 2 was much more "significant" than was replication 1. In attempting to see if this is so, we considered the Bayesian approach to study of bias within the framework of the aspirin example.

EXAMPLE 2. For simplicity in the aspiring example, we reduce consideration to

- θ = true difference in heart attack rates between aspirin and placebo populations multiplied by 1000;
- Y = difference in observed heart attack rates between aspirin and placebo groups in original study multiplied by 1000;
- X_i = difference in observed heart attack rates between aspirin and placebo groups in Replication i multiplied by 1000.

We assume that the replication studies are extremely well designed and implemented, so that one is very confident that the X_i have mean θ. Using normal approximations for convenience, the data can be summarized as

$$X_1 \sim N(x_1 \mid \theta, 4.82), \quad X_2 \sim N(x_2 \mid \theta, 3.63)$$

with actual observations $x_1 = 7.704$ and $x_2 = 13.07$.

Consider now the bias issue. We assume that the original experiment is somewhat suspect in this regard, and we will model bias by defining the mean of Y to be

$$\eta = \theta + \beta,$$

where β is the unknown bias. Then the data in the original experiment can be summarized by

$$Y \sim N(y \mid \eta, 1.54),$$

with the actual observation being $y = 7.707$.

Bayesian analysis requires specification of a prior distribution, $\pi(\beta)$, for the suspected amount of bias. Of particular interest then are the posterior distribution of β, assuming replication i has been performed, given by

$$\pi(\beta \mid y, x_i) \propto \pi(\beta)\exp\left\{-\frac{1}{2(1.54^2 + \sigma_i^2)}[\beta - (y - x_i)]^2\right\},$$

where σ_i^2 is the variance (4.82 or 3.63) from replication i; and the posterior probability of H_0, given by

$$P(H_0 \mid y, x_i) = \int_{-\infty}^{\infty} \Phi\left(-\frac{\sigma_i}{1.54\sqrt{\sigma_i^2 + 1.54^2}}(y - \beta) - \frac{1.54}{\sigma_i\sqrt{\sigma_i^2 + 1.54^2}}x_i\right)\pi(\beta \mid y, x_i)\, d\beta.$$

Recall that our goal here was to see if Bayesian analysis can reproduce the intuition that the original experiment could be trusted if replication 1 had been done, while it could not be trusted (in spite of its much larger sample size) had replication 2 been performed. Establishing this requires finding a prior distribution $\pi(\beta)$ for which $\pi(\beta \mid y, x_1)$ has little effect on $P(H_0 \mid y, x_1)$, but $\pi(\beta \mid y, x_2)$ has a large effect on $P(H_0 \mid y, x_2)$. To achieve the first objective, $\pi(\beta)$ must be tightly concentrated near zero. To achieve the second, $\pi(\beta)$ must be such that large $|y - x_2|$, which suggests presence of a large bias, can result in a substantial shift of posterior mass for β away from zero.

A sensible candidate for the prior density $\pi(\beta)$ is the Cauchy (0, V) density

$$\pi_V(\beta) = \frac{1}{\pi V[1 + (\theta/V)^2]}.$$

Flat-tailed densities, such as this, are well known to have the property that when discordant data is observed (e.g., when ($|y - x_2|$ is large), substantial mass shifts away from the prior center towards the likelihood center. It is easy to see that a normal prior for β can not have the desired behavior.

Our first surprise in consideration of these priors was how small V needed to be chosen in order for $P(H_0 | y, x_1)$ to be unaffected by the bias. For instance, even with $V = 1.54/100$ (recall that 1.54 was the standard deviation of Y from the original experiment), computation yields $P(H_0 | y, x_1) = 4.3 \times 10^{-5}$, compared with the P-value (and posterior probability from the original experiment assuming no bias) of 2.8×10^{-7}. There is a clear lesson here; even very small suspicions of bias can drastically alter a small P-value. Note that replication 1 is very consistent with the presence of no bias, and so the posterior distribution for the bias remains tightly concentrated near zero; for instance, the mean of the posterior for β is then 7.2×10^{-6}, and the standard deviation is 0.25.

When we turned attention to replication 2, we found that it did not seriously change the prior perceptions of bias. Examination quickly revealed the reason; even the maximum likelihood estimate of the bias is no more than 1.4 standard deviations from zero, which is not enough to change strong prior beliefs. We, therefore, considered a third experiment, defined in Table 1. Transforming to approximate normality, as before, yields

$$X_3 \sim N(x_3 | \theta, 3.48),$$

with $x_3 = 22.72$ being the actual observation. The maximum likelihood estimate of bias is now 3.95 standard deviations from zero, so there is potential for a substantial change in opinion about the bias.

Sure enough, computation when $V = 1.54/100$ yields that $E[\beta | y, x_3] = -4.9$ with (posterior) standard deviation equal to 6.62, which is a dramatic shift from prior opinion (that β is Cauchy (0, 1.54/100)). The effect of this is to essentially ignore the original experiment in overall assessments of evidence. For instance, $P(H_0 | y, x_3) = 3.81 \times 10^{-11}$, which is very close to $P(H_0 | x_3) = 3.29 \times 10^{-11}$. Note that, if β were set equal to zero, the overall posterior probability of H_0 (and P-value) would be 2.62×10^{-13}.

Thus Bayesian reasoning can reproduce the intuition that replication which indicates bias can cast considerable doubt on the original experiment, while replication which provides no evidence of bias leaves evidence from the original experiment intact. Such behavior seems only obtainable, however, with flat-tailed priors for bias (such as the Cauchy) that are very concentrated (in comparison with the experimental standard deviation) near zero.

TABLE 1
Frequency of heart attacks in replication 3

	Yes	No
Aspirin	5	2309
Placebo	54	2116

3. P-VALUES OR BAYES FACTORS?

Parapsychology experiments usually consider testing of H_0: No parapsychological effect exists. Such null hypotheses are often realistically represented as point nulls (see Berger and Delampady, 1987, for the reason that care must be taken in such representation), in which case it is known that there is a large difference between P-values and posterior probabilities (see Berger and Delampady, 1987, for review). The article by Jefferys (1990) dramatically illustrates this, showing that a very small P-value can actually correspond to evidence for H_0 when considered from a Bayesian perspective. (This is very related to the famous "Jeffreys" paradox.) The argument in favor of the Bayesian approach here is very strong, since it can be shown that the conflict holds for virtually any sensible prior distribution; a Bayesian answer can be wrong if the prior information turns out to be inaccurate, but a Bayesian answer that holds for all sensible priors is unassailable.

Since P-values simply cannot be viewed as meaningful in these situations, we found it of interest to reconsider the example in Section 5 from a Bayes factor perspective. We considered only analysis of the overall totals, that is, $x = 122$ successes out of $n = 355$ trials. Assuming a simple Bernoulli trial model with success probability θ, the goal is to test $H_0: \theta = 1/4$ versus $H_1: \theta \neq 1/4$.

To determine the Bayes factor here, one must specify $g(\theta)$, the conditional prior density on H_1. Consider choosing g to be uniform and symmetric, that is,

$$G_r(\theta) = \begin{cases} \dfrac{1}{2r}, & \text{for } \dfrac{1}{4} - r \le \theta \le \dfrac{1}{4} + r, \\ 0, & \text{otherwise}. \end{cases}$$

Crudely, r could be considered to be the maximum change in success probability that one would expect given that ESP exists. Also, these distributions are the "extreme points" over the class of symmetric unimodal conditional densities, so answers that hold over this class are also representative of answers over a much larger class. Note that here $r \le 0.25$ (because $0 \le \theta \le 1$); for the given data the $\theta > 0.5$ are essentially irrelevant, but if it were deemed important to take them into account one could use the more sophisticated binomial analysis in Berger and Delampady (1987).

For g_r, the Bayes factor of H_1 to H_0, which is to be interpreted as the relative odds for the hypotheses provided by the data, is given by

$$B(r) = \frac{(1/(2r)) \int_{.25-r}^{.25+r} \theta^{122}(1-\theta)^{355-122}\, d\theta}{(1/4)^{122}(1-1/4)^{355-122}}$$

$$\cong \frac{1}{2r}(63.13) \cdot \left[\Phi\left(\frac{r-.0937}{.0252}\right) + \Phi\left(\frac{-(r+.0937)}{.0252}\right)\right].$$

This is graphed in Figure 1.

The P-value for this problem was 0.00005, indicating overwhelming evidence against H_0 from a classical perspective. In contrast to the situation studied by Jefferys (1990), the Bayes factor here does not completely reverse the conclusion, showing that there are very reasonable values of r for which the evidence against H_0 is moderately strong, for example 100/1 or 200/1. Of course, this evidence is probably not of sufficient strength to overcome strong prior opinions against H_0 (one

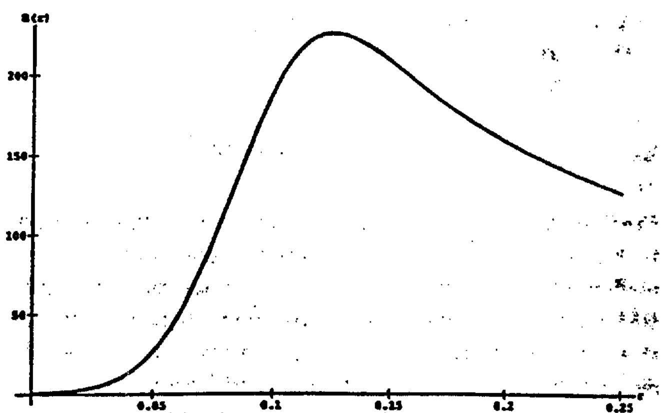

FIG. 1. *The Bayes factor of H_1 to H_0 as a function of r, the maximum change in success probability that is expected given that ESP exists, for the ganzfeld experiment.*

obtains final posterior odds by multiplying prior odds by the Bayes factor). To properly assess strength of evidence, we feel that such Bayes factor computations should become standard in parapsychology.

As mentioned by Professor Utts, Bayesian methods have additional potential in situations such as this, by allowing unrealistic models of iid trials to be replaced by hierarchical models reflecting differing abilities among subjects.

ACKNOWLEDGMENTS

M. J. Bayarri's research was supported in part by the Spanish Ministry of Education and Science under DGICYT Grant BE91-038, while visiting Purdue University. James Berger's research was supported by NSF Grant DMS-89-23071.

Comment

Ree Dawson

This paper offers readers interested in statistical science multiple views of the controversial history of parapsychology and how statistics has contributed to its development. It first provides an account of how both design and inferential aspects of statistics have been pivotal issues in evaluating the outcomes of experiments that study psi abilities. It then emphasizes how the idea of science as replication has been key in this field in which results have not been conclusive or consistent and thus meta-analysis has been at the heart of the literature in parapsychology. The author not only reviews past debate on how to interpret repeated psi studies, but also provides very detailed information on the Honorton–Hyman argument, a nice illustration of the challenges of resolving such de-

Ree Dawson is Senior Statistician, New England Biomedical Research Foundation, and Statistical Consultant, RFE/RL Research Institute. Her mailing address is 177 Morrison Avenue, Somerville, Massachusetts 02144.

bate. This debate is also a good example of how statistical criticism can be part of the scientific process and lead to better experiments and, in general, better science.

The remainder of the paper addresses technical issues of meta-analysis, drawing upon recent research in parapsychology for an in-depth application. Through a series of examples, the author presents a convincing argument that power issues cannot be overlooked in successive replications and that comparison of effect sizes provides a richer alternative to the dichotomous measure inherent in the use of p-values. This is particularly relevant when the potential effect size is small and resources are limited, as seems to be the case for psi studies.

The concluding section briefly mentions Bayesian techniques. As noted by the author, Bayes (or empirical Bayes) methodology seems to make sense for research in parapsychology. This discussion examines possible Bayesian approaches to meta-analysis in this field.

BAYES MODELS FOR PARAPSYCHOLOGY

The notion of repeatability maps well into the Bayesian set-up in which experiments, viewed as a random sample from some superpopulation of experiments, are assumed to be exchangeable. When subjects can also be viewed as an approximately random sample from some population, it is appropriate to pool them across experiments. Otherwise, analyses that partially pool information according to experimental heterogeneity need to be considered. Empirical and hierarchical Bayes methods offer a flexible modeling framework for such analyses, relying on empirical or subjective sources to determine the degree of pooling. These richer methods can be particularly useful to meta-analysis of experiments in parapsychology conducted under potentially diverse conditions.

For the recent ganzfeld series, assuming them to be independent binomially distributed as discussed in Section 5, the data can be summed (pooled) across series to estimate a common hit rate. Honorton et al. (1990) assessed the homogeneity of effects across the 11 series using a chi-square test that compares individual effect sizes to the weighted mean effect. The chi-square statistic $\chi^2_{10} = 16.25$, not statistically significant ($p = 0.093$), largely reflects the contribution of the last "special" series (contributes 9.2 units to the χ^2_{10} value), and to a lesser extent the novice series with a negative effect (contributes 2.5 units). The outlier series can be dropped from the analysis to provide a more conservative estimate of the presence of psi effects for this data (this result is reported in Section 5). For the remaining 10 series, the chi-square value $\chi^2_9 = 7.01$ strongly favors homogeneity, although more than one-third of its value is due to the novice series (number 4 in Table 1). This pattern points to the potential usefulness of a richer model to accommodate series that may be distinct from the others. For the earlier ganzfeld data analyzed by Honorton (1985b), the appeal of a Bayes or other model that recognizes the heterogeneity across studies is clear cut: $\chi^2_{23} = 56.6$, $p = 0.0001$, where only those studies with common chance hit rate have been included (see Table 2).

Historic reliance on voting-count approaches to determine the presence of psi effects makes it natural to consider Bayes models that focus on the ensemble of experimental effects from parapsychological studies, rather than individual estimates. Recent work in parapsychology that compares effect sizes across studies, rather than estimating separate study effects, reinforces the need to examine this type of model. Louis (1984) develops Bayes and empirical Bayes methods for problems that consider the ensemble of parameter values to be the primary goal, for example, multiple comparisons. For the simple compound normal model, $Y_i \sim N(\theta_i, 1)$, $\theta_i \sim N(\mu, \tau^2)$, the standard Bayes estimates (posterior means)

$$\theta_i^* = \mu + D(Y_i - \mu) \quad \text{and} \quad D = \frac{\tau^2}{1 + \tau^2}$$

where the θ_i represent experimental effects of interest, are modified approximately to

$$\theta_i^l \approx \mu + \sqrt{D}\,(Y_i - \mu)$$

when an ensemble loss function is assumed. The new estimates adjust the shrinkage factor D so that their sample mean and variance match the posterior expectation and variance of the θ's. Similar results are obtained when the model is gener-

TABLE 1
Recent ganzfeld series

Series type	N Trials	Hit rate	Y_i	σ_i
Pilot	22	0.36	−0.58	0.44
Pilot	9	0.33	−0.71	0.71
Pilot	36	0.28	−0.94	0.37
Novice	50	0.24	−1.15	0.33
Novice	50	0.36	−0.58	0.30
Novice	50	0.30	−0.85	0.31
Novice	50	0.36	−0.58	0.30
Novice	6	0.67	0.71	0.87
Experienced	7	0.43	−0.28	0.76
Experienced	50	0.30	−0.85	0.31
Experienced	25	0.64	0.58	0.42
Overall	355	0.34		

TABLE 2
Earlier ganzfeld studies

N Trials	Hit rate	Y_i	σ_i
32	0.44	−0.24	0.36
7	0.86	1.82	1.09
30	0.43	−0.28	0.37
30	0.23	−1.21	0.43
20	0.10	−2.20	0.75
10	0.90	2.20	1.05
10	0.40	−0.41	0.65
28	0.29	−0.90	0.42
10	0.40	−0.41	0.65
20	0.35	−0.62	0.47
26	0.31	−0.80	0.42
20	0.45	−0.20	0.45
20	0.45	−0.20	0.45
30	0.53	0.12	0.37
36	0.33	−0.71	0.35
32	0.28	−0.94	0.39
40	0.28	−0.94	0.35
26	0.46	−0.16	0.39
20	0.60	0.41	0.46
100	0.41	−0.36	0.20
40	0.33	−0.71	0.34
27	0.41	−0.36	0.39
60	0.45	−0.20	0.26
48	0.21	−1.33	0.35
722	.38		

alized to the case of unequal variances, $Y_i \sim N(\theta_i, \sigma_i^2)$.

For the above model, the fraction of θ_i^l above (or below) a cut point C is a consistent estimate of the fraction of $\theta_i > C$ (or $\theta_i < C$). Thus, the use of ensemble, rather than component-wise, loss can help detect when individual effects are above a specified threshold by chance. For the meta-analysis of ganzfeld experiments, the observed binomial proportions transformed on the logit (or arcsin$\sqrt{}$) scale can be modeled in this framework. Letting d_i and m_i denote the number of direct hits and misses respectively for the ith experiment, and p_i as the corresponding population proportion of direct hits, the Y_i are the observed logits

$$Y_i = \log(d_i / m_i)$$

and σ_i^2, estimated by maximum likelihood as $1/d_i + 1/m_i$, is the variance of Y_i conditional on $\theta_i = \text{logit}(p_i)$. The threshold logit (0.25) = 1.10 can be used to identify the number of experiments for which the proportion of direct hits exceeds that expected by chance.

Table 1 shows Y_i and σ_i for the 11 ganzfeld series. All but one of the series are well above the threshold; Y_4 marginally falls below −1.10. Any shrinkage toward a common hit rate will lead to an estimate, θ_4^* or θ_4^l, above the threshold. The use of ensemble loss (with its consistency property) provides more convincing support that all $\theta_i > -1.10$, although posterior estimates of uncertainty are needed to fully calibrate this. For the earlier ganzfeld data in Table 2, ensemble loss can similarly be used to determine the number of studies with $\theta_i < -1.10$ and specifically whether the negative effects of studies 4 and 24 ($Y_4 = -1.21$ and $Y_{24} = -1.33$) occurred as a result of chance fluctuation.

Features of the ganzfeld data in Section 5, such as the outlier series, suggest that further elaboration of the basic Bayesian set-up may be necessary for some meta-analyses in parapsychology. Hierarchical models provide a natural framework to specify these elaborations and explore how results change with the prior specification. This type of sensitivity analysis can expose whether conclusions are closely tied to prior beliefs, as observed by Jeffreys for RNG data (see Section 7). Quantifying the influence of model components deemed to be more subjective or less certain is important to broad acceptance of results as evidence of psi performance (or lack thereof).

Consider the initial model commonly used for Bayesian analysis of discrete data:

$$Y_i \mid p_i, n_i \sim B(p_i, n_i),$$
$$\theta_i \sim N(\mu, \tau^2), \quad \theta_i = \text{logit}(p_i),$$

with noninformative priors assumed for μ and τ^2 (e.g., log τ locally uniform). The distinctiveness of the last "special" series and, in general, the different types of series (pilot versus formal, novice versus experienced) raises the question of whether the experimental effects follow a normal distribution. Weighted normal plots (Ryan and Dempster, 1984) can be used to graphically diagnose the adequacy of second-stage normality (see Dempster, Selwyn and Weeks, 1983, for examples with binary response and normal superpopulation).

Alternatively, if nonnormality is suspected, the model can be revised to include some sort of heavy-tailed prior to accommodate possibly outlying series or studies. West (1985) incorporates additional scale parameters, one for each component of the model (experiment), that flexibly adapt to a typical θ_i and discount their influence on posterior estimates, thus avoiding under- or over-shrinkage due to such θ_i. For example, the second stage can specify the prior as a scale mixture of normals:

$$\theta_i \sim N(\mu, \tau^2 \gamma_i^{-1}),$$
$$k\gamma_i \sim \chi_k^2,$$
$$\upsilon\tau^{-2} \sim \chi_\upsilon^2.$$

This approach for the prior is similar to others for maximum likelihood estimation that modify the sampling error distribution to yield estimates that are "robust" against outlying observations.

Like its maximum likelihood counterparts, in addition to the robust effect estimates θ_i^*, the Bayes model provides (posterior) scale estimates γ_i^*. These can be interpreted as the weight given to the data for each θ_i in the analysis and are useful to diagnosing which model components (series or studies) are unusual and how they influence the shrinkage. When more complex groupings among the θ_i are suspected, for example, bimodal distribution of studies from different sites or experimenters, other mixture specifications can be used to further relax the shrinkage toward a common value.

For the 11 ganzfeld series, the last "outlier" series, quite distinct from the others (hit rate = 0.64), is moderately precise ($N = 25$). Omitting it from the analysis causes the overall hit rate to drop from 0.344 to 0.321. The scale mixture model is a compromise between these two values (on the logit scale), discounting the influence of series 11 on the estimated posterior common hit rate used for shrinkage. The scale factor γ_{11}^*, an indication of how separate θ_{11} is from the other parameters, also causes θ_{11}^* to be shrunk less toward the common hit rate than other, more homogeneous θ_i, giving more weight to individual information for that series (see West, 1985). The heterogeneity of the earlier ganzfeld data is more pronounced, and studies are taken from a variety of sources over time. For these data, the γ_i^* can be used to explore atypical studies (e.g., study 6, with hit rate = 0.90, contributes more than 25% to the χ_{23}^2 value for homogeneity) and groupings among effects, as well as protect the analysis from misspecification of second-stage normality.

Variation among ganzfeld series or studies and the degree to which pooling or shrinking is appropriate can be investigated further by considering a range of priors for τ^2. If the marginal likelihood of τ^2 dominates the prior specification, then results should not vary as the prior for τ^2 is varied. Otherwise, it is important to identify the degree to which subjective information about interexperimental variability influences the conclusions. This sensitivity analysis is a Bayesian enrichment of the simpler test of homogeneity directed toward determining whether or not complete pooling is appropriate.

To assess how well heterogeneity among historical control groups is determined by the data. Dempster, Selwyn and Weeks (1983) propose three priors for τ^2 in the logistic-normal model. The prior distributions range from strongly favoring individual estimates, $p(\tau^2)d\tau \propto \tau^{-1}$, to the uniform reference prior $p(\tau^2)d\tau \propto \tau^{-2}$, flat on the log τ scale, to strongly favoring complete pooling, $p(\tau^2)d\tau \propto \tau^{-3}$ (the latter forcing complete pooling for the compound normal model; see Morris, 1983). For their two examples, the results (estimates of linear treatment effects) are largely insensitive to variation in the prior distribution, but the number of studies in each example was large (70 and 19 studies available for pooling). For the 11 ganzfeld series, τ^2 may be less well determined by the data. The posterior estimate of τ^2 and its sensitivity to $p(\tau^2)d\tau$ will also depend on whether individual scale parameters are incorporated into the model. Discounting the influence of the last series will both shift the marginal likelihood toward smaller values of τ^2 and concentrate it more in that region.

The issue of objective assessment of experiment results is one that extends well beyond the field of parapsychology, and this paper provides insight into issues surrounding the analysis and interpretation of small effects from related studies. Bayes methods can contribute to such meta-analyses in two ways. They permit experimental and subjective evidence to be formally combined to determine the presence or absence of effects that are not clear cut or controversial (e.g., psi abilities). They can also help uncover sources and degree of uncertainty in the scientific conclusions.

Comment

Persi Diaconis

In my experience, parapsychologists use statistics extremely carefully. The plethora of widely significant p-values in the many thousands of published parapsychological studies must give us pause for thought. Either something spooky is going on, or it is possible for a field to exist on error and artifact for over 100 years. The present paper offers a useful review by an expert and a glimpse at some tantalizing new studies.

My reaction is that the studies are crucially flawed. Since my reasons are somewhat unusual, I will try to spell them out.

I have found it impossible to usefully judge what actually went on in a parapsychology trial from their published record. Time after time, skeptics have gone to watch trials and found subtle and not-so-subtle errors. Since the field has so far failed to produce a replicable phenomena, it seems to me that any trial that asks us to take its findings seriously should include full participation by qualified skeptics. Without a magician and/or knowledgeable psychologist skilled at running experiments with human subjects, I don't think a serious effort is being made.

I recognize that this is an unorthodox set of requirements. In fact, one cannot judge what "really goes on" in studies in most areas, and it is impossible to demand wide replicability in others. Finally, defining "qualified skeptic" is difficult. In defense, most areas have many easily replicable experiments and many have their findings explained and connected by unifying theories. It simply seems clear that when making claims at such extraordinary variance with our daily experience, claims that have been made and washed away so often in the past, such extraordinary measures are mandatory before one has the right to ask outsiders to spend their time in review. The papers cited in Section 5 do not actively involve qualified skeptics, and I do not feel they have earned the right to our serious attention.

The points I have made above are not new. Many appear in the present article. This does not diminish their utility nor applicability to the most recent studies.

Parapsychology is worth serious study. First, there may be something there, and I marvel at the patience and drive of people like Jessica Utts and Ray Hyman. Second, if it is wrong, it offers a truly alarming massive case study of how statistics can mislead and be misused. Third, it offers marvelous combinatorial and inferential problems. Chung, Diaconis, Graham and Mallows (1981), Diaconis and Graham (1981) and Samaniego and Utts (1983) offer examples not cited in the text. Finally, our budding statistics students are fascinated by its claims; the present paper gives a responsible overview providing background for a spectacular classroom presentation.

Persi Diaconis is Professor of Mathematics at Harvard University, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

Comment: Parapsychology—On the Margins of Science?

Joel B. Greenhouse

Professor Utts reviews and synthesizes a large body of experimental literature as well as the scientific controversy involved in the attempt to establish the existence of paranormal phenomena. The organization and clarity of her presentation are noteworthy. Although I do not believe that this paper will necessarily change anyone's views regarding the existence of paranormal phenomena, it does raise very interesting questions about the process by which new ideas are either accepted or rejected by the scientific community. As students of science, we believe that scientific discovery

Joel B. Greenhouse is Associate Professor of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890.

advances methodically and objectively through the accumulation of knowledge (or the rejection of false knowledge) derived from the implementation of the scientific method. But, as we will see, there is more to the acceptance of new scientific discoveries than the systematic accumulation and evaluation of facts. The recognition that there is a social process involved with the acceptance or rejection of scientific knowledge has been the subject of study of sociologists for some time. The scientific community's rejection of the existence of paranormal phenomena is an excellent case study of this process (Allison, 1979; Collins and Pinch, 1979).

Implicit in Professor Utts' presentation and paramount to the acceptance of parapsychology as a legitimate science are the description and documentation of the professionalization of the field of parapsychology. It is true that many researchers in the field have university appointments; there are organized professional societies for the advancement of parapsychology; there are journals with rigorous standards for published research; the field has received funding from federal agencies; and parapsychology has received recognition from other professional societies, such as the IMS and the American Association for the Advancement of Science (Collins and Pinch, 1979). Nevertheless, most readers of *Statistical Science* would agree that parapsychology is not accepted as part of orthodox science and is considered by most of the scientific community to be on the margins of science, at best (Allison, 1979; Collins and Pinch, 1979). Why is this the case? Professor Utts believes that it is because people have not examined the data. She states that "Strong beliefs tend to be resistant to change even in the face of data, and many people, scientists included, seem to have made up their minds on the question without examining any empirical data at all."

The history of science is replete with examples of resistance by the established scientific community to new discoveries. A challenging problem for science is to understand the process by which a new theory or discovery becomes accepted by the community of scientists and, likewise, to characterize the nature of the resistance to new ideas. Barber (1961) suggests that there are many different sources of resistance to scientific discovery. In 1900, for example, Karl Pearson met resistance to his use of statistics in applications to biological problems, illustrating a source of resistance due to the use of a particular methodology. The Royal Society informed Pearson that future papers submitted to the Society for publication must keep the mathematics separate from the biological applications.

Another obvious source of resistance to new scientific ideas, and the one referred to by Professor Utts above, is the prevailing substantive beliefs and theories held by scientists at any given time. Barber offers the opposition to Copernicus and his heliocentric theory and to Mendel's theory of genetic inheritance as examples of how, because of preconceived ideas, theories and values, scientists are not as open-minded to new advances as one might think they should be. It was R. A. Fisher who said that each generation seems to have found in Mendel's paper only what it expected to find and ignored what did not conform to its own expectations (Fisher, 1936).

Pearson's response to the antimathematical prejudice expressed by the Royal Society was to establish with Galton's support a new journal, *Biometrika*, to encourage the use of mathematics in biology. Galton (1901) wrote an article for the first issue of the journal, explaining the need for this new voice of "mutual encouragement and support" for mathematics in biology and saying that "a new science cannot depend on a welcome from the followers of the older ones, and [therefore]... it is advisable to establish a special Journal for Biometry." Lavoisier understood the role of preconceived beliefs as a source of resistance when he wrote in 1785,

> I do not expect my ideas to be adopted all at once. The human mind gets creased into a way of seeing things. Those who have envisaged nature according to a certain point of view during much of their career, rise only with difficulty to new ideas. (Barber, 1961.)

I suspect that this paper by Professor Utts synthesizing the accumulation of research results supporting the existence of paranormal phenomena will continue to be received with skepticism by the orthodox scientific community "even after examining the data." In part, this resistance is due to the popular perception of the association between parapsychology and the occult (Allison, 1979) and due to the continued suspicion and documentation of fraud in parapsychology (Diaconis, 1978). An additional and important source of resistance to the evidence presented by Professor Utts, however, is the lack of a model to explain the phenomena. Psychic phenomena are unexplainable by any current scientific theory and, furthermore, directly contradict the laws of physics. Acceptance of psi implies the rejection of a large body of accumulated evidence explaining the physical and biological world as we know it. Thus, even though the effect size for a relationship between aspirin and the prevention of heart attacks is three times smaller than the effect size observed in the ganzfeld data base, it is the existence of a biological mechanism to explain the effectiveness of aspirin that accounts, in part, for acceptance of this relationship.

In evaluating the evidence in favor of the existence of paranormal phenomena, it is necessary to consider alternative explanations or hypotheses for the results and, as noted by Cornfield (1959), "If important alternative hypotheses are compatible with available evidence, then the question is unsettled, even if the evidence is experimental" (see also Platt, 1964). Many of the experimental results reported by Professor Utts need to be considered in the context of explanations other than the existence of paranormal phenomena. Consider the following examples:

(1) In the various psi experiments that Professor Utts discusses, the null hypothesis is a simple chance model. However, as noted by Diaconis (1978) in a critique of parapsychological research, "In complex, badly controlled experiments simple chance models cannot be seriously considered as tenable explanations: hence, rejection of such models is not of particular interest." Diaconis shows that the underlying probabilistic model in many of these experiments (even those that are well controlled) is much more complicated than chance.

(2) The role that experimenter expectancy plays in the reporting and interpreting of results cannot be underestimated. Rosenthal (1966), based on a meta-analysis of the effects of experimenters' expectancies on the results of their research, found that experimenters tended to get the results they expected to get. Clearly this is an important potential confounder in parapsychological research. Professor Utts comments on a debate between Honorton and Hyman, parapsychologist and critic, respectively, regarding evidence for psi abilities, and, although not necessarily a result of experimenter expectancy, describes how "...each analyzed the results of all known psi ganzfeld experiments to date, and reached strikingly different conclusions."

(3) What is an acceptable response in these experiments? What constitutes a direct hit? What if the response is close, who decides whether or not that constitutes a hit (see (2) above)? In an example of a response of a Receiver in an automated ganzfeld procedure, Professor Utts describes the "dream-like quality of the mentation." Someone must evaluate these stream-of-consciousness responses to determine what is a hit. An important methodological question is: How sensitive are the results to different definitions of a hit?

(4) In describing the results of different meta-analyses, Professor Utts is careful to raise questions about the role of publication bias. Publication bias or "the file-drawer problem" arises when only statistically significant findings get published, while statistically nonsignificant studies sit unreported in investigators' file drawers. Typically, Rosenthal's method (1979) is used to calculate the "fail-safe N," that is, the number of unreported studies that would have to be sitting in file-drawers in order to negate the significant effect. Iyengar and Greenhouse (1988) describe a modification of Rosenthal's method, however, that gives a fail-safe N that is often an order of magnitude smaller than Rosenthal's method, suggesting that the sensitivity of the results of meta-analyses of psi experiments to unpublished negative studies is greater than is currently believed.

Even if parapsychology is thought to be on the margins of science by the scientific community, parapsychologists should not be held to a different standard of evidence to support their findings than orthodox scientists, but like other scientists they must be concerned with spurious effects and the effects of extraneous variables. The experimental results summarized by Professor Utts appear to be sensitive to the effect of alternative hypotheses like the ones described above. Sensitivity analyses, which question, for example, how large of an effect due to experimenter expectancy there would have to be to account for the effect sizes being reported in the psi experiments, are not addressed here. Again, the ability to account for and eliminate the role of alternative hypotheses in explaining the observed relationship between aspirin and the prevention of heart attacks is another reason for the acceptance of these results.

A major new technology discussed by Professor Utts in synthesizing the experimental parapsychology literature is meta-analysis. Until recently, the quantitative review and synthesis of a research literature, that is, meta-analysis, was considered by many to be a questionable research tool (Wachter, 1988). Resistance by statisticians to meta-analysis is interesting because, historically, many prominent statisticians found the combining of information from independent studies to be an important and useful methodology (see, e.g., Fisher, 1932; Cochran, 1954; Mosteller and Bush, 1954; Mantel and Haenszel, 1959). Perhaps the more recent skepticism about meta-analysis is because of its use as a tool to advance discoveries that themselves were the objects of resistance, such as the efficacy of psychotherapy (Smith and Glass, 1977) and now the existence of paranormal phenomena. It is an interesting problem for the history of science to explore why and when in the development of a of a discipline it turns to meta-analysis to answer research questions or to resolve controversy (e.g., Greenhouse et al., 1990).

One argument for combining information from different studies is that a more powerful result can be obtained than from a single study. This objective is implicit in the use of meta-analysis in parapsychology and is the force behind Professor Utts' paper. The issue is that by combining many small studies consisting of small effects there is a gain in power to find an overall statistically significant effect. It is true that the meta-analyses reported by Professor Utts find extremely small p-values, but the estimate of the overall effect size is still small. As noted earlier, because of the small magnitude of the overall effect size, the possibility that other extraneous variables might account for the relationship remains.

Professor Utts, however, also illustrates the use of meta-analysis to investigate how studies differ and to characterize the influence of difficult covariates or moderating variables on the combined estimate of effect size. For example, she compares the mean effect size of studies where subjects were selected on the basis of good past performance to studies where the subjects were unselected, and she compares the mean effect size of studies with feedback to studies without feedback. To me, this latter use of meta-analysis highlights the more valuable and important contribution of the methodology. Specifically, the value of quantitative methods for research synthesis is in assessing the potential effects of study characteristics and to quantify the sources of heterogeneity in a research domain, that is, to study systematically the effects of extraneous variables. Tom Chalmers and his group at Harvard have used meta-analysis in just this way not only to advance the understanding of the effectiveness of medical therapies but also to study the characteristics of good research in medicine, in particular, the randomized controlled clinical trial. (See Mosteller and Chalmers, 1991, for a review of this work.)

Professor Utts should be congratulated for her courage in contributing her time and statistical expertise to a field struggling on the margins of science, and for her skill in synthesizing a large body of experimental literature. I have found her paper to be quite stimulating, raising many interesting issues about how science progresses or does not progress.

ACKNOWLEDGMENT

This work was supported in part by MHCRC grant MH30915 and MH15758 from the National Institute of Mental Health, and CA54852 from the National Cancer Institute. I would like to acknowledge stimulating discussions with Professors Larry Hedges, Michael Meyer, Ingram Olkin, Teddy Seidenfeld and Larry Wasserman, and thank them for their patience and encouragement while preparing this discussion.

Comment

Ray Hyman

Utts concludes that "there is an anomaly that needs explanation." She bases this conclusion on the ganzfeld experiments and four meta-analyses of parapsychological studies. She argues that both Honorton and Rosenthal have successfully refuted my critique of the ganzfeld experiments. The meta-analyses apparently show effects that cannot be explained away by unreported experiments nor over-analysis of the data. Furthermore, effect size does not correlate with the rated quality of the experiment.

Ray Hyman is Professor of Psychology, University of Oregon, Eugene, Oregon 97403.

Neither time nor space is available to respond in detail to her argument. Instead, I will point to some of my concerns. I will do so by focusing on those parts of Utts' discussion that involve me. Understandably, I disagree with her assertions that both Honorton and Rosenthal successfully refuted my criticisms of the ganzfeld experiments.

Her treatment of both the ganzfeld debate and the National Research Council's report suggests that Utts has relied on second-hand reports of the data. Some of her statements are simply inaccurate. Others suggest that she has not carefully read what my critics and I have written. This remoteness from the actual experiments and details of the arguments may partially account for her optimistic assessment of the results. Her paper takes the reported data at face value and focuses on the statistical interpretation of these data.

Both the statistical interpretation of the results of an individual experiment and of the results of a meta-analysis are based on a model of an ideal world. In this ideal world, effect sizes have a tractable and known distribution and the points in the sample space are independent samples from a coherent population. The appropriateness of any statistical application in a given context is an empirical matter. That is why such issues as the adequacy of randomization, the non-independence of experiments in a meta-analysis and the over-analysis of data are central to the debate. The optimistic conclusions from the meta-analyses assume that the effect sizes are unbiased estimates from independent experiments and have nicely behaved distributional properties.

Before my detailed assessment of all the available ganzfeld experiments through 1981, I accepted the assertions by parapsychologists that their experiments were of high quality in terms of statistical and experimental methodology. I was surprised to find that the ganzfeld experiments, widely heralded as the best exemplar of a successful research program in parapsychology, were characterized by obvious possibilities for sensory leakage, inadequate randomization, over-analysis and other departures from parapsychology's own professed standards. One response was to argue that I had exaggerated the number of flaws. But even internal critics agreed that the rate of defects in the ganzfeld data base was too high.

The other response, implicit in Utts' discussion of the ganzfeld experiments and the meta-analyses, was to admit the existence of the flaws but to deny their importance. The parapsychologists doing the meta-analysis would rate each experiment for quality on one or more attributes. Then, if the null hypothesis of no correlation between effect size and quality were upheld, the investigators concluded that the results could not be attributed to defects in methodology.

This retrospective sanctification using statistical controls to compensate for inadequate experimental controls has many problems. The quality ratings are not blind. As the differences between myself and Honorton reveal, such ratings are highly subjective. Although I tried my best to restrict my ratings to what I thought were objective and easily codeable indicators, my quality ratings provide a different picture than do those of Honorton. Honorton, I am sure, believes he was just as objective in assigning his ratings as I believe I was.

Another problem is the number of different properties that are rated. Honorton's ratings of quality omitted many attributes that I included in my ratings. Even in those cases where we used the same indicators to make our assessments, we differed because of our scaling. For example, on adequacy of randomization I used a simple dichotomy. Either the experimenter clearly indicated using an appropriate randomization procedure or he did not. Honorton converted this to a trichotomous scale. He distinguished between a clearly inadequate procedure such as hand-shuffling and failure to report how the randomization was done. He then assigned the lowest rating to failure to describe the randomization. In his scheme, clearly inadequate randomization was of higher quality than failure to describe the procedure. Although we agreed on which experiments had adequate randomization, inadequate randomization or inadequate documentation, the different ways these were ordered produced important differences between us in how randomization related to effect size. These are just some of the reasons why the finding of no correlation between effect size and rated quality does not justify concluding that the observed flaws had no effect.

I will now consider some of Utts' assertions and hope that I can go into more detail in another forum. Utts discusses the conclusions of the National Research Council's Committee on Techniques for the Enhancement of Human Performance. I was chairperson of that committee's subcommittee on paranormal phenomena. She wrongly states that we restricted our evaluation only to significant studies. I do not know how she got such an impression since we based our analysis on meta-analyses whenever these were available. The two major inputs for the committee's evaluation were a lengthy evaluation of contemporary parapsychology experiments by John Palmer and an independent assessment of these experiments by James Alcock. Our sponsors, the Army Research Institute had commissioned the report from the parapsychologist John Palmer. They specifically asked our committee to provide a second opinion from a non-parapsychological perspective. They were most interested in the experiments on remote viewing and random number generators. We decided to add the ganzfeld experiments. Alcock was instructed, in making his evaluation, to restrict himself to the same experiments in these categories that Palmer had chosen. In this way, the experiments we evaluated, which included both significant and nonsignificant ones, were, in effect, selected for us by a prominent parapsychologist.

Utts mistakenly asserts that my subcommittee on parapsychology commissioned Harris and Rosenthal to evaluate parapsychology experiments for us. Harris and Rosenthal were commissioned by our evaluation subcommittee to write a paper on evaluation issues, especially those related to experimenter effects. On their own initiative, Harris and Rosenthal surveyed a number of data bases to illustrate the application of methodological procedures such as meta-analysis. As one illustration, they included a meta-analysis of the subsample of ganzfeld experiments used by Honorton in his rebuttal to my critique.

Because Harris and Rosenthal did not themselves do a first-hand evaluation of the ganzfeld experiments, and because they used Honorton's ratings for their illustration, I did not refer to their analysis when I wrote my draft for the chapter on the paranormal. Rosenthal told me, in a letter, that he had arbitrarily used Honorton's ratings rather than mine because they were the most recent available. I assumed that Harris and Rosenthal were using Honorton's sample and ratings to illustrate meta-analytic procedures. I did not believe they were making a substantive contribution to the debate.

Only after the committee's complete report was in the hands of the editors did someone become concerned that Harris and Rosenthal had come to a conclusion on the ganzfeld experiments different from the committee. Apparently one or more committee members contacted Rosenthal and asked him to explain why he and Harris were dissenting.

Because some committee members believed that we should deal with this apparent discrepancy, I contacted Rosenthal and pointed out if he had used my ratings with *the very same analysis* he had applied to Honorton's ratings, he would have reached a conclusion opposite to what Harris and he had asserted. I did this, not to suggest my ratings were necessarily more trustworthy than Honorton's, but to point out how fragile any conclusions were based on this small and limited sample. Indeed, the data were so lacking in robustness that the difference between my rating and Honorton's rating of one investigator (Sargent) on one attribute (randomization) sufficed to reverse the conclusions Harris and Rosenthal made about the correlation between quality and effect size.

Harris and Rosenthal responded by adding a footnote to their paper. In this footnote, they reported an analysis using my ratings rather than Honorton's. This analysis, they concluded, still supported the null hypothesis of no correlation between quality and effect size. They used 6 of my 12 dichotomous ratings of flaws as predictors and the z score and effect size as criterion variables in both multiple regression and canonical correlation analyses. They reported an "adjusted" canonical correlation between criterion variables and flaws of "only" 0.46. A true correlation of this magnitude would be impressive given the nature and split of the dichotomous variables. But, because it was not statistically significant, Harris and Rosenthal concluded that there was no relationship between quality and effect size. A canonical correlation on this sample of 28 nonindependent cases, of course, has virtually no chance of being significant, even if it were of much greater magnitude.

What this amounts to is that the alleged contradictory conclusions of Harris and Rosenthal are based on a meta-analysis that supports Honorton's position when Honorton's ratings are used and supports my position when my ratings are used. Nothing substantive comes from this, and it is redundant with what Honorton and I have already published. Harris and Rosenthal's footnote adds nothing because it supports the null hypothesis with a statistical test that has no power against a reasonably sized alternative. It is ironic that Utts, after emphasizing the importance of considering statistical power, places so much reliance on the outcome of a powerless test.

(I should add that the recurrent charge that the NRC committee completely ignored Harris and Rosenthal's conclusions is not strictly correct. I wrote a response to the Harris and Rosenthal paper that was included in the same supplementary volume that contains their commissioned paper.)

Utts' discussion of the ganzfeld debate, as I have indicated, also shows unfamiliarity with details. She cites my factor analysis and Saunders' critique as if these somehow jeopardized the conclusions I drew. Again, the matter is too complex to discuss adequately in this forum. The "factor analysis" she is talking about is discussed in a few pages of my critique. I introduced it as a convenient way to summarize my conclusions, *none of which depended on this analysis*. I agree with what Saunders has to say about the limitations of factor analysis in this context. Unfortunately, Saunders bases his criticism on wrong assumptions about what I did and why I did it. His dismissal of the results as "meaningless" is based on mistaken algebra. I included as dummy variables five experimenters in the factor analysis. Because an experimenter can only appear on one variable, this necessarily forces the average intercorrelation among the experimenter variables to be negative. Saunders falsely asserts that this negative correlation must be -1. If he were correct, this would make the results meaningless. But he could be correct only if there were just two investigators and that each one accounted for 50% of the experiments. In my case, as I made sure to check ahead of time, the use of five experimenters, each of whom contributed only a few studies to the data base, produced a mildly negative intercorrelation of −0.147. To make sure even that small correlation did not distort the results, I did the factor analysis with and without the dummy variables. The same factors were obtained in both cases.

However, I do not wish to defend this factor analysis. None of my conclusions depend on it. I would agree with any editor who insisted that I omit it from the paper on the grounds of redundancy. I am discussing it here as another example that suggests that Utts is not familiar with some relevant details in literature she discusses.

CONCLUSIONS

Utts may be correct. There may indeed be an anomaly in the parapsychological findings. Anomalies may also exist in non-parapsychological domains. The question is when is an anomaly worth taking seriously. The anomaly that Utts has in mind, if it exists, can be described only as a departure from a generalized statistical model. From the evidence she presents, we might conclude that we are dealing with a variety of different anomalies instead of one coherent phenomenon. Clearly, the reported effect sizes for the experiments with random number generators are orders of magnitude lower than those for the ganzfeld experiments. Even within the same experimental domain, the effect sizes do not come from the same population. The effects sizes obtained by Jahn are much smaller than those obtained by Schmidt with similar experiments on random number generators. In the ganzfeld experiments, experimenters differ significantly in the effect sizes each obtains.

This problem of what effect sizes are and what they are measuring points to a problem for parapsychologists. In other fields of science such as astronomy, an "anomaly" is a very precisely specified departure from a well-established substantive theory. When Leverrier discovered Neptune by studying the perturbations in the orbit of Uranus, he was able to characterize the anomaly as a very precise departure of a specific kind from the orbit expected on the basis of Newtonian mechanics. He knew exactly what he had to account for.

The "anomaly" or "anomalies" that Utts talks about are different. We do not know what it is that we are asked to account for other than something that sometimes produces nonchance departures from a statistical model, whose appropriateness is itself open to question.

The case rests on a handful of meta-analyses that suggest effect sizes different from zero and uncorrelated with some non-blindly determined indices of quality. For a variety of reasons, these retrospective attempts to find evidence for paranormal phenomena are problematical. At best, they should provide the basis for parapsychologists designing prospective studies in which they can specify, in advance, the complete sample space and the critical region. When they get to the point where they can specify this along with some boundary conditions and make some reasonable predictions, then they will have demonstrated something worthy of our attention.

In this context, I agree with Utts that Honorton's recent report of his automated ganzfeld experiments is a step in the right direction. He used the ganzfeld meta-analyses and the criticisms of the existing data base to design better experiments and make some predictions. Although he and Utts believe that the findings of meaningful effect sizes in the dynamic targets and a lack of a nonzero effect size in the static targets are somehow consistent with previous ganzfeld results, I disagree. I believe the static targets are closer in spirit to the original data base. But this is a minor criticism.

Honorton's experiments have produced intriguing results. If, as Utts suggests, independent laboratories can produce similar results with the same relationships and with the same attention to rigorous methodology, then parapsychology may indeed have finally captured its elusive quarry. Of course, on several previous occasions in its century-plus history, parapsychology has felt it was on the threshold of a breakthrough. The breakthrough never materialized. We will have to patiently wait to see if the current situation is any different.

Comment

Robert L. Morris

Experimental sciences by their nature have found it relatively easy to deal with simple closed systems. When they come to study more complex, open systems, however, they have more difficulty in generating testable models, must rely more on multivariate approaches, have more diversity from experiment to experiment (and thus more difficulty in constructing replication attempts), have more noise in the data, and more difficulty in constructing a linkage between concept and measurement. Data gatherers and other researchers are more likely to be part of the system themselves. Examples include ecology, economics, social psychology and parapsychology. Parapsychology can be regarded as the study of apparent new means of communication, or transfer of influence, between organism and environment. Any observer attempting to decide whether or not such psychic communication has taken place is one of several elements in a complex open system composed of an indefinite number of interactive features. The system can be modeled, as has been done elsewhere (e.g., Morris, 1986) such as to organise our understanding of how observers can be misled by themselves, or by deliberate frauds. Parapsychologists designing experimental studies must take extreme care to ensure that the elements in the experimental system do not interact in unanticipated ways to produce artifact or encourage fraudulent procedures. When researchers follow up the findings of others, they must ensure that the new experimental system sufficiently resembles the earlier one, regarding its important components and their potential interactions. Specifying sufficient resemblance is more difficult in complex and open systems, and in areas of research using novel methodologies.

As a result, parapsychology and other such areas may well profit from the application of modern meta-analysis, and meta-analytic methods may in turn profit from being given a good stiff workout by controversial data bases, as suggested by Jessica Utts in her article. Parapsychology would appear to gain from meta-analytic techniques, in at least three important areas.

First, in assessing the question of replication rate, the new focus on effect size and confidence intervals rather than arbitrarily chosen significance levels seems to indicate much greater consistency in the findings than has previously been claimed.

Second, when one codes the individual studies for flaws and relates flaw abundance with effect size, there appears to be little correlation for all but one data base. This contradicts the frequent assertion that parapsychological results disappear when methodology is tightened. Additional evidence on this point is the series of studies by Honorton and associates using an automated ganzfeld procedure, apparently better conducted than any of the previous research, which nevertheless obtained an effect size very similar to that of the earlier more diverse data base.

Third, meta-analysis allows researchers to look at moderator variables, to build a clearer picture of the conditions that appear to produce the strongest effects. Research in any real scientific discipline must be cumulative, with later researchers building on the work of those who preceded them. If our earlier successes and failures have meaning, they should help us obtain increasingly consistent, clearer results. If psychic ability exists and is sufficiently stable that it can be manifest in controlled experimental studies, then moderator variables should be present in groups of studies that would indicate conditions most favourable and least favourable to the production of large effect sizes. From the analyses presented by Utts, for instance, it seems evident that group studies tend to produce poor results and, however convenient it may be to conduct them, future researchers should apparently focus much more on individual testing. When doing ganzfeld studies, it appears best to work with dynamic rather than static target material and with experienced participants rather than novices. If such results are valid, then future researchers who wish to get strong results now have a better idea of what procedures to select to increase the likelihood of so doing, what elements in the experimental system seem most relevant. The proportion of studies obtaining positive results should therefore increase.

However, the situation may be more complex than the somewhat ideal version painted above. As noted earlier, meta-analysis may learn from parapsychology as well as vice versa. Parapsychological data may well give meta-analytic techniques a good workout and will certainly pose some challenges. None of the cited meta-analyses, as described above,

Robert L. Morris occupies the Koestler Chair of Parapsychology in the Department of Psychology at the University of Edinburgh, 7 George Square, Ed-

evaluator. Certainly none of them cited any correlation values between evaluators, and the correlations between judges of research quality in other social sciences tend to be "at best around .50," according to Hunter and Schmidt (1990, page 497). Although Honorton and Hyman reported a relatively high correlation of 0.77 between themselves, they were each doing their own study and their flaw analyses did reach somewhat different conclusions, as noted by Utts. Other than Hyman, the evaluators cited by Utts tend to be positively oriented toward parapsychology; roughly speaking, all evaluators doing flaw analyses found what they might hope to find, with the exception of the PK dice data base. Were evaluators blind as to study outcome when coding flaws? No comment is made on this aspect. The above studies need to be replicated, with multiple (and blind) evaluators and reported indices of evaluator agreement. Ideally, evaluator attitude should be assessed and taken into account as well. A study with all hostile evaluators may report very high evaluator correlations, yet be a less valid study than one that employs a range of evaluators and reports lower correlations among evaluators.

But what constitutes a replication of a meta-analysis? As with experimental replications, it may be important to distinguish between exact and conceptual replications. In the former, a replicator would attempt to match all salient features of the initial analysis, from the selection of reports to the coding of features to the statistical tests employed, such as to verify that the stated original protocol had been followed faithfully and that a similar outcome results. For conceptual replication, replicators would take the stated outcome of the meta-analysis and attempt their own independent analysis, with their own initial report selection criteria, coding criteria and strategy for statistical testing, to see if similar conclusions resulted. Conceptual replication allows more room for bias and resultant debate when findings differ, but when results are similar they can be assumed to have more legitimacy. Given the strong and surprising (for many) conclusions reached in the meta-analysis reported by Utts, it is quite likely that others with strong views on parapsychology will attempt to replicate, hoping for clear confirmation or disconfirmation. The diversity of methods they are likely to employ and the resultant debates should provide a good opportunity for airing the many conceptual problems still present in meta-analysis. If results differ on moderator variables, there can come to be empirical resolution of the differences as further results unfold. With regard to flaw analysis, such analyses have already focused attention in ganzfeld research on the abundance of existing faults and how to avoid them. If results are as strong under well-controlled conditions as under sloppy ones, then additional research such as that done by Honorton and associates under tight conditions should continue to produce positive results.

In addition to the replication issue, there are some other problems that need to be addressed. So far, the assessment of moderator variables has been univariate, whereas a multivariate approach would seem more likely to produce a clearer picture. Moderator variables may covary, with each other or with flaws. For instance, in the dice data higher effect sizes were found for flawed studies and for studies with selected subjects. Did studies using special subjects use weaker procedures?

Given the importance attached to effect size and incorporating estimates of effect size in designing studies for power, we must be careful not to assume that effect size is independent of number of trials or subjects unless we have empirical reason to do so. Effect size may decrease with larger N if experimenters are stressed or bored towards the end of a long study or if there are too many trials to be conducted within a short period of time and subjects are given less time to absorb their instructions or to complete their tasks. On one occasion there is presentation of an estimated "true average effect size," (0.18 rather than 0.28) without also presenting an estimate of effect size dispersal. Future investigators should have some sense of how the likelihood that they will obtain a hit rate of 1/3 (where 1/4 is expected) will vary in accordance with conditions.

There are a few additional quibbles with particular points. In Utts' example experiment with Professor A versus Professor B, sex of professor is a possible confounding variable. When Honorton omitted studies that did not report direct hits as a measure, he may have biased his sample. Were there studies omitted that could have reported direct hits but declined to do so, conceivably because they looked at that measure, saw no results and dropped it? This objection is only with regard to the initial meta-analysis and is not relevant for the later series of studies which all used direct hits. In Honorton's meta-analysis of forced-choice precognition experiments, the comparison variables of feedback delay and time interval to target selection appear to be confounded. Studies delaying target selection cannot provide trial by trial feedback, for instance. Also, I am unsure about using an approximation to Cohen's h for assessing the effect size for the aspirin study. There would appear to be a very striking effect, with the aspirin condition heart attack rate only 55% that of the rate for the placebo condition. How was the expected proportion of misses estimated; perhaps Cohen's *h* greatly underestimates effect size when very low probability events (less than 1 in 50 for heart attack in the placebo condition and less than 1 in a 100 for aspirin) are involved. I'm not a statistician and thus don't know if there is a relevant literature on this point.

The above objections should not detract from the overall value of the Utts survey. The findings she reports will need to be replicated; but even as is, they provide a challenge to some of the cherished arguments of counteradvocates, yet also challenge serious researchers to use these findings effectively as guidelines for future studies.

Comment

Frederick Mosteller

Dr. Utts's discussion stimulates me to offer some comments that bear on her topic but do not, in the main, fall into an agree-disagree mode. My references refer to her bibliography.

Let me recommend J. Edgar Coover's work to statisticians who would like to read about a pretty sequence of experiments developed and executed well before Fisher's book on experimental design appeared. Most of the standard kinds of ESP experiments (though not the ganzfeld) are carried out and reported in this 1917 book. Coover even began looking into the amount of information contained in cues such as whispers. He also worked at exposing mediums. I found the book most impressive. As Utts says in her article, the question of significance level was a puzzling one, and one we still cannot solve even though some fields seem to have standardized on 0.05.

When Feller's comments on Stuart and Greenwood's sampling experiments came out in the first edition of his book, I was surprised. Feller devotes a problem to the results of generating 25 symbols from the set a, b, c, d and e (page 45, first edition) using random numbers with 0 and 1 corresponding to a, 2 and 3 to b, etc. He asks the student to find out how often the 25 produce 5 of each symbol. He asks the student to check the results using random number tables. The answer seems to be about 1 chance in 500. In a footnote Feller then says "They [random numbers] are occasionally extraordinarily obliging: c.f. J. A. Greenwood and E. E. Stuart, Review of Dr. Feller's Critique, *Journal of Parapsychology*, vol. 4 (1940), pp. 298–319, in particular p. 306." The 25 symbols of 5 kinds, 5 of each, correspond to the cards in a parapsychology deck.

The point of page 306 is that Greenwood and Stuart on that page claim to have generated two random orders of such a deck using Tippett's table of random numbers. Apparently Feller thought that it would have taken them a long time to do it. If one assumes that Feller's way of generating a random shuffle is required, then it would indeed be unreasonable to suppose that the experiments could be carried out quickly. I wondered then whether Feller thought this was the only way to produce a random order to such a deck of cards. If you happen to know how to shuffle a deck efficiently using random numbers, it is hard to believe that others do not know. I decided to test it out and so I proposed to a class of 90 people in mathematical statistics that we find a way of using random numbers to shuffle a deck of cards. Although they were familiar with random numbers, they could not come up with a way of doing it, nor did anyone after class come in with a workable idea though several students made proposals. I concluded that inventing such a shuffling technique was a hard problem and that maybe Feller just did not know how at the time of writing the footnote. My face-to-face attempts to verify this failed because his response was evasive. I also recall Feller speaking at a scientific meeting where someone had complained about mistakes in published papers. He said essentially that we won't have any literature if mistakes are disallowed and further claimed that he always had mistakes in his own papers, hard as he tried to avoid them. It was fun to hear him speak.

Although I find Utts's discussion of replication engaging as a problem in human perception, I do always feel that people should not be expected to carry out difficult mathematical exercises in their head, off the cuff, without computers, textbooks or advisors. The kind of problem treated requires

Frederick Mosteller is Roger I. Lee Professor of Mathematical Statistics, Emeritus, at Harvard University and Director of the Technology Assessment Group in the Harvard School of Public Health. His mailing address is Department of Statistics, Harvard University, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.

after a careful analysis is completed, there can be vigorous reasonable arguments about the appropriateness of the formulation and its analysis. These investigations leave me reinforced with the belief that people cannot do hard mathematical problems in their heads, rather than with an attitude toward or against ESP investigations.

When I first became aware of the work of Rhine and others, the concept seemed to me to be very important and I asked a psychologist friend why more psychologists didn't study this field. He responded that there were too many ways to do these experiments in a poorly controlled manner. At the time, I had just discovered that when viewed with light coming from a certain angle, I could read the backs of the cards of my parapsychology deck as clearly as the faces. While preparing these remarks in 1991, I found a note on page 305 of volume 1 of *The Journal of Parapsychology* (1937) indicating that imperfections in the cards precluded their use in unscreened situations, but that improvements were on the way. Thus I sympathize with Utts's conclusion that much is to be gained by studying how to carry out such work well. If there is no ESP, then we want to be able to carry out null experiments and get no effect, otherwise we cannot put much belief in work on small effects in non-ESP situations. If there is ESP, that is exciting. However, thus far it does not look as if it will replace the telephone.

Rejoinder

Jessica Utts

I would like to thank this distinguished group of discussants for their thought-provoking contributions. They have raised many interesting and diverse issues. Certain points, such as Professor Mosteller's enlightening account of Feller's position, require no further comment. Other points indicate the need for clarification and elaboration of my original material. Issues raised by Professors Diaconis and Hyman and subsequent conversations with Robert Rosenthal and Charles Honorton have led me to consider the topic of "Satisfying the Skeptics." Since the conclusion in my paper was not that psychic phenomena have been proved, but rather that there is an anomalous effect that needs to be explained, comments by several of the discussants led me to address the question "Should Psi Research be Ignored by the Scientific Community?" Finally, each of the discussants addressed replication and modeling issues. The last part of my rejoinder comments on some of these ideas and discusses them in the context of parapsychology.

CLARIFICATION AND ELABORATION

Since my paper was a survey of hundreds of experiments and many published reports, I could obviously not provide all of the details to accompany this overview. However, there were details lacking in my paper that have led to legitimate questions and misunderstandings from several of the discussants. In this section, I address specific points raised by Professors Diaconis, Greenhouse, Hyman and Morris, by either clarifying my original statements or by adding more information from the original reports.

Points Raised by Diaconis

Diaconis raised the point that qualified skeptics and magicians should be active participants in parapsychology experiments. I will discuss this general concept in the next section, but elaborate here on the steps that were taken in this regard for the autoganzfeld experiments described in Section 5 of my paper. As reported by Honorton et al. (1990):

> Two experts on the simulation of psi ability have examined the autoganzfeld system and protocol. Ford Kross has been a professional mentalist [a magician who simulates psychic abilities] for over 20 years... Mr. Kross has provided us with the following statement: "In my professional capacity as a mentalist, I have reviewed Psychophysical Research Laboratories' automated ganzfeld system and found it to provide excellent security against deception by subjects." We have received similar comments from Daryl Bem, Professor of Psychology at Cornell University. Professor Bem is well known for his research in social and personality psychology. He is also a member of the Psychic Entertainers Association and has performed for many years as a mentalist. He vis-

Approved For Release 2003/04/18 : CIA-RDP96-00789R002700010001-1

ited PRL for several days and was a subject in Series 101" [pages 134–135].

Honorton has also informed me (personal communication, July 25, 1991) that several self-proclaimed skeptics have visited his laboratory and received demonstrations of the autoganzfeld procedure and that no one expressed any concern with the security arrangements.

This may not completely satisfy Professor Diaconis' objections, but it does indicate a serious effort on the part of the researchers to involve such people. Further, the original publication of the research in Section 5 followed the reporting criteria established by Hyman and Honorton (1986), thus providing much more detail for the reader than the earlier published records to which Professor Diaconis alludes.

Points Raised by Greenhouse

Greenhouse enumerated four items that offer alternative explanations for the observed anomalous effects. Three of these (items 2–4) will be addressed in this section by elaborating on the details provided in my paper. His item 1 will be addressed in a later section.

Item 2 on his list questioned the role of experimenter expectancy effects as a potential confounder in parapsychological research. While the expectations of the experimenter may influence the *reporting* of results, the ganzfeld experiments (as well as other psi experiments) are conducted in such a way that experimenter expectancy cannot account for the results themselves. Rosenthal, who Greenhouse cites as the expert in this area, addressed this in his background paper for the National Research Council (Harris and Rosenthal, 1988a) and concluded that the ganzfeld studies were adequately controlled in this regard. He also visited the autoganzfeld laboratory and was given a demonstration of that procedure.

Greenhouse's item 3, the question of what constitutes a direct hit, was addressed in my paper but perhaps needs elaboration. Although free-response experiments do generate substantial amounts of subjective data, the statistical analysis requires that the results for each trial be condensed into a single measure of whether or not a direct hit was achieved. This is done by presenting four choices to a judge (who of course does not know the correct answer) and asking the judge to decide which of the four best matches the subject's response. If the judge picks the target, a direct hit has occurred.

It is true that different judges may differ on their opinions of whether or not there has been a direct hit on any given trial, but in all cases the statistical question is the same. Under the null hypothesis, since the target is randomly selected from the four possibilities presented, the probability of a direct hit is 0.25 regardless of who does the judging. Thus, the observed anomalous effects cannot be explained by assuming there was an over-optimistic judge.

If Professor Greenhouse is suggesting that the source of judging may be a moderating variable that determines the magnitude of the demonstrated anomalous effect, I agree. The parapsychologists have considered this issue in the context of whether or not subjects should serve as judges for their own sessions, with differing opinions in different laboratories. This is an example of an area that has been suggested for further research.

Finally, Greenhouse raised the question of the accuracy of the file-drawer estimates used in the reported meta-analyses. I agree that it is instructive to examine the file-drawer estimate using more than one model. As an example, consider the 39 studies from the direct hit and autoganzfeld data bases. Rosenthal's fail-safe N estimates that there would have to be 371 studies in the file-drawer to account for the results. In contrast, the method proposed by Iyengar and Greenhouse gives a file-drawer estimate of 258 studies. Even this estimate is unrealistically large for a discipline with as few researchers as parapsychology. Given that the average number of trials per experiment is 30, this would represent almost 8000 unreported trials, and at least that many hours of work.

There are pros and cons to any method of estimating the number of unreported studies, and the actual practices of the discipline in question should be taken into account. Recognizing publication bias as an issue, the Parapsychological Association has had an official policy since 1975 against the selective reporting of positive results. Of the original ganzfeld studies reported in Section 4 of my paper, less than half were significant, and it is a matter of record that there are many nonsignificant studies and "failed replications" published in all areas of psi research. Further, the autoganzfeld database reported in Section 5 has no file-drawer. Given the publication practices and the size of the field, the proposed file-drawer cannot account for the observed effects.

Points Raised by Hyman

One of my goals in writing this paper was to present a fair account of recent work and debate in parapsychology. Thus, I was disturbed that Hyman, who has devoted much of his career to the study of parapsychology, and who had first-hand knowledge of the original published reports, believed that some of my statements were inaccurate and indicated that I had not carefully read the reports. I will address some of his specific objections and show that, except where noted, the accuracy of my original statements can be verified by further elaboration and clarification, with due apology for whatever necessary details were lacking in my original report.

Most of our points of disagreement concern the National Academy of Sciences (National Research Council) report *Enhancing Human Performance* (Druckman and Swets, 1988). This report evaluated several controversial areas, including parapsychology. Professor Hyman chaired the Parapsychology Subcommittee. Several background papers were commissioned to accompany this report, available from the "Publication on Demand Program" of the National Academy Press. One of the papers was written by Harris and Rosenthal, and entitled "Human Performance Research: An Overview."

Professor Hyman alleged that "Utts mistakenly asserts that my subcommittee on parapsychology commissioned Harris and Rosenthal to evaluate parapsychology experiments for us...." I cannot find a statement in my paper that asserts that Harris and Rosenthal were commissioned by the subcommittee, nor can I find a statement that asserts that they were asked to evaluate parapsychology experiments. Nonetheless, I believe our substantive disagreement results from the fact that the work by Harris and Rosenthal was written in two parts, both of which I referenced in my paper. They were written several months apart, but published together, and each had its own history.

The first part (Harris and Rosenthal, 1988a) is the one to which I referred with the words "Rosenthal was commissioned by the National Academy of Sciences to prepare a background paper to accompany its 1988 report on parapsychology" (p. 372). According to Rosenthal (personal communication, July 23, 1991) he was asked to prepare a background paper to address evaluation issues and experimenter effects to accompany the report in five specific areas of research, including parapsychology.

The second part was a "Postscript" to the commissioned paper (Harris and Rosenthal, 1988b), and this is the one to which I referred on page 371 as "requested by Hyman in his capacity as Chair of the National Academy of Sciences' Subcommittee on Parapsychology." (It is probably this wording that led Professor Hyman to his erroneous allegation.) The postscript began with the words "We have been asked to respond to a letter from Ray Hyman, chair of the subcommittee on parapsychology, in which he raises questions about the presence and consequence of methodological flaws in the ganzfeld studies...."

In reference to this postscript, I stand corrected on a technical point, because Hyman himself did not request the response to his own letter. As noted by Palmer, Honorton and Utts (1989), the postscript was added because:

> At one stage of the process, John Swets, Chair of the Committee, actually phoned Rosenthal and asked him to withdraw the parapsychology section of his [commissioned] paper. When Rosenthal declined, Swets and Druckman then requested that Rosenthal respond to criticisms that Hyman had included in a July 30, 1987 letter to Rosenthal [page 38].

A related issue on which I would like to elaborate concerns the correlation between flaws and success in the original ganzfeld data base. Hyman has misunderstood both my position and that of Harris and Rosenthal. He believes that I implicitly denied the importance of the flaws, so I will make my position explicit. I do not think there is any evidence that the experimental results were due to the identified flaws. The flaw analysis was clearly useful for delineating acceptable criteria for future experiments. Several experiments were conducted using those criteria. The results were similar to the original experiments. I believe that this indicates an anomaly in need of an explanation.

In discussing the paper and postscript by Harris and Rosenthal, Hyman stated that "The alleged contradictory conclusions [to the National Research Council report] of Harris and Rosenthal are based on a meta-analysis that supports Honorton's position when Honorton's [flaw] ratings are used and supports my position when my ratings are used." He believes that Harris and Rosenthal (and I) failed to see this point because the low power of the test associated with their analysis was not taken into account.

The analysis in question was based on a canonical correlation between flaw ratings and measures of successful outcome for the ganzfeld studies. The canonical correlation was 0.46, a value Hyman finds to be impressive. What he has failed to take into account however, is that a canonical correlation gives only the *magnitude* of the relationship, and not the *direction*. A careful reading of Harris and Rosenthal (1988b) reveals that their analysis actually *contradicted* the idea that the flaws could account for the successful ganzfeld results, since "Interestingly, three of the six flaw variables correlated positively with the flaw canonical variable and with the outcome canonical variable but three correlated *negatively*" (page 2, italics added). Rosenthal (personal communication, July 23, 1991) verified that this was indeed the point he was trying to make. Readers who are interested in drawing their own conclusions from first-hand analyses can find Hyman's original flaw codings in an Appendix to his paper (Hyman, 1985, pages 44–49).

Finally, in my paper, I stated that the parapsychology chapter of the National Research Council report critically evaluated statistically significant experiments, but not those that were nonsignificant. Professor Hyman "does not know how [I] got such an impression," so I will clarify by outlining some of the material reviewed in that report. There were surveys of three major areas of psi research: remote viewing (a particular type of free-response experiment), experiments with random number generators, and the ganzfeld experiments. As an example of where I got the impression that they evaluated only significant studies, consider the section on remote viewing. It began by referencing a published list of 28 studies. Fifteen of these were immediately discounted, since "only 13 ... were published under refereed auspices" (Druckman and Swets, 1988, page 179). Four more were then dismissed, since "Of the 13 scientifically reported experiments, 9 are classified as successful" (page 179). The report continued by discussing these nine experiments, never again mentioning any of the remaining 19 studies. The other sections of the report placed similar emphasis on significant studies. I did not think this was a valid statistical method for surveying a large body of research.

Minor Point Raised by Morris

The final clarification I would like to offer concerns the minor point raised by Professor Morris, that "When Honorton omitted studies that did not report direct hits as a measure, he may have biased his sample." This possibility was explicitly addressed by Honorton (1985, page 59). He examined what would happen if *z*-scores of zero were inserted for the 10 studies for which the number of direct hits was not measured, but could have been. He found that even with this conservative scenario, the combined *z*-score only dropped from 6.60 to 5.67.

SATISFYING THE SKEPTICS

Parapsychology is probably the only scientific discipline for which there is an organization of skeptics trying to discredit its work. The Committee for the Scientific Investigation of Claims of the Paranormal (CSICOP) was established in 1976 by philosopher Paul Kurtz and sociologist Marcello Truzzi when "Kurtz became convinced that the time was ripe for a more active crusade against parapsychology and other pseudo-scientists" (Pinch and Collins, 1984, page 527). Truzzi resigned from the organization the next year (as did Professor Diaconis) "because of what he saw as the growing danger of the committee's excessive negative zeal at the expense of responsible scholarship" (Collins and Pinch, 1982, page 84). In an advertising brochure for their publication *The Skeptical Inquirer*, CSICOP made clear its belief that paranormal phenomena are worthy of scientific attention only to the extent that scientists can fight the growing interest in them. Part of the text of the brochure read: "Why the sudden explosion of interest, even among some otherwise sensible people, in all sorts of paranormal 'happenings'? ... Ten years ago, scientists started to fight back. They set up an organization—The Committee for the Scientific Investigation of Claims of the Paranormal."

During the six years that I have been working with parapsychologists, they have repeatedly expressed their frustration with the unwillingness of the skeptics to specify what would constitute acceptable evidence, or even to delineate criteria for an acceptable experiment. The Hyman and Honorton Joint Communiqué was seen as the first major step in that direction, especially since Hyman was the Chair of the Parapsychology Subcommittee of CSICOP.

Hyman and Honorton (1986) devoted eight pages to "Recommendations for Future Psi Experiments," carefully outlining details for how the experiments should be conducted and reported. Honorton and his colleagues then conducted several hundred trials using these specific criteria and found essentially the same effect sizes as in earlier work for both the overall effect and effects with moderator variables taken into account. I would expect Professor Hyman to be very interested in the results of these experiments he helped to create. While he did acknowledge that they "have produced intriguing results," it is both surprising and disappointing that he spent only a scant two paragraphs at the end of his discussion on these results.

Instead, Hyman seems to be proposing yet another set of requirements to be satisfied before parapsychology should be taken seriously. It is difficult to sort out what those requirements should be from his account: "[They should] specify, in advance, the complete sample space and the critical region. When they get to the point where they can specify this along with some boundary conditions and make some reasonable predictions, then they will have demonstrated something worthy of our attention."

Diaconis believes that psi experiments do not deserve serious attention unless they actively involve skeptics. Presumably, he is concerned with subject or experimenter fraud, or with improperly controlled experiments. There are numerous documented cases of fraud and trickery in purported psychic phenomena. Some of these were observed by Diaconis and reported in his article in *Science*. Such cases have mainly been revealed when investigators attempted to verify the claims of individual psychic practitioners in quasi-experimental or uncontrolled conditions. These instances have received considerable attention, probably because the claims are so sensational, the fraud is so easy to detect by a skilled observer and they are an easy target for skeptics looking for a way to discredit psychic phenomena. As noted by Hansen (1990), "Parapsychology has long been tainted by the fraudulent behavior of a few of those claiming psychic abilities" (page 25).

Control against deception by subjects in the laboratory has been discussed extensively in the parapsychological literature (see, e.g., Morris, 1986, and Hansen, 1990). Properly designed experiments should preclude the possibility of such fraud. Hyman and Honorton (1986, page 355) explicitly discussed precautions to be taken in the ganzfeld experiments, all of which were followed in the autoganzfeld experiments. Further the controlled laboratory experiments discussed in my paper usually used a large number of subjects, a situation that minimizes the possibility that the results were due to fraud on the part of a few subjects. As for the possibility of experimenter fraud, it is of course an issue in all areas of science. There have been a few such instances in parapsychology, but since parapsychologists tend to be aware of this possibility, they were generally detected and exposed by insiders in the field.

It is not clear whether or not Diaconis is suggesting that a magician or "qualified skeptic" needs to be present at all times during a laboratory experiment. I believe that it would be more productive for such consultation to occur during the design phase, and during the implementation of some pilot sessions. This is essentially what was done for the autoganzfeld experiments, in which Professor Hyman, a skeptic as well as an accomplished magician, participated in the specification of design criteria, and mentalists Bem and Kross observed experimental sessions. Bem is also a well-respected experimental psychologist.

While I believe that the skeptics, particularly some of the more knowledgeable members of CSICOP, have served a useful role in helping to improve experiments, their counter-advocacy stance is counterproductive. If they are truly interested in resolving the question of whether or not psi abilities exist, I would expect them to encourage evaluation and experimentation by unbiased, skilled experimenters. Instead, they seem to be trying to discourage such interest by providing a moving target of requirements that must be satisfied first.

SHOULD PSI RESEARCH BE IGNORED BY THE SCIENTIFIC COMMUNITY?

In the conclusion of my paper, I argued that the scientific community should pay more attention to the experimental results in parapsychology. I was not suggesting that the accumulated evidence constitutes proof of psi abilities, but rather that it indicates that there is indeed an anomalous effect that needs an explanation. Greenhouse noted that my paper will not necessarily change anyone's view about the existence of paranormal phenomena, an observation with which I agree. However, I hope it will change some views about the importance of further investigation.

Mosteller and Diaconis both acknowledged that there are reasons for statisticians to be interested in studying the anomalous effects, regardless of whether or not psi is real. As noted by Mosteller, "If there is no ESP, then we want to be able to carry out null experiments and get no effect, otherwise we cannot put much belief in work on small effects in non-ESP situations." Diaconis concluded that "Parapsychology is worthy of serious study" partly because "If it is wrong, it offers a truly alarming massive case study of how statistics can mislead and be misused."

Greenhouse noted several sociological reasons for the resistance of the scientific community to accepting parapsychological phenomena. One of these is that they directly contradict the laws of physics. However, this assertion is not uniformly accepted by physicists (see, e.g., Oteri, 1975), and some of the leading parapsychological researchers hold Ph.D.s in physics.

Another reason cited by Greenhouse, and supported by Hyman, is that psychic phenomena are currently unexplainable by a unified scientific theory. But that is precisely the reason for more intensive investigation. The history of science and medicine is replete with examples where empirical departures from expectation led to important findings or theoretical models. For example, the causal connection between cigarette smoking and lung cancer was established only after years of statistical studies, resulting from the observation by one physician that his lung cancer patients who smoked did not recover at the same rate as those who did not. There are many medications in common use for which there is still no medical explanation for their observed therapeutic effectiveness, but that does not prohibit their use.

There are also examples where a coherent theory of a phenomenon was impossible because the requisite background information was missing. For instance, the current theory of endorphins as an explanation for the success of acupuncture would have been impossible before the discovery of endorphins in the 1970s.

Mosteller's observation that ESP will not replace the telephone leads to the question of whether or not psi abilities are of any use even if they do exist, since the effects are relatively small. Again, a look at history is instructive. For example, in 1938 *Fortune Magazine* reported that "At present, few scientists foresee any serious or practical use for atomic energy."

Greenhouse implied that I think parapsychology is not accepted by more of the scientific community only because they have not examined the data, but this misses the main point I was trying to make. The point is that individual scientists are willing to express an opinion without any reference to data. The interesting sociological question is why they are so resistant to examining the data. One of the major reasons is undoubtedly the perception identified by Greenhouse that there is some connection between parapsychology and the occult, or worse, religious beliefs. Since religion is clearly not in the realm of science, the very thought that parapsychology might be a science leads to what psychologists call "cognitive dissonance." As noted by Griffin (1988), "People feel unpleasantly aroused when two cognitions are dissonant—when they contradict one another" (page 33). Griffin continued by observing that there are also external reasons for scientists to discount the evidence, since "It is generally easier to be a skeptic in the face of novel evidence; skeptics may be overly conservative, but they are rarely held up to ridicule" (page 34).

In summary, while it may be safer and more consonant with their beliefs for individual scientists to ignore the observed anomalous effects, the scientific community should be concerned with finding an explanation. The explanations proposed by Greenhouse and others are simply not tenable.

REPLICATION AND MODELING

Parapsychology is one of the few areas where a point null hypothesis makes some sense. We can specify what should happen if there is no such thing as ESP by using simple binomial models, either to find p-values or Bayes factors. As noted by Mosteller, if there is no ESP, or other nonstatistical explanation for an effect, we should be able to carry out null experiments and get no effect. Otherwise, we should be worried about using these simple models for other applications.

Greenhouse, in his first alternative explanation for the results, questioned the use of these simple models, but his criticisms do not seem relevant to the experiments discussed in Section 5 of my paper. The experiments to which he referred were either poorly controlled, in which case no statistical analysis could be valid, or were specifically designed to incorporate trial by trial feedback in such a way that the analysis needed to account for the added information. Models and analyses for such experiments can be found in the references given at the end of Diaconis' discussion.

For the remainder of this discussion, I will confine myself to models appropriate for experiments such as the autoganzfeld described in Section 5. It is this scenario for which Bayarri and Berger computed Bayes factors, and for which Dawson discussed possible Bayesian models.

If ESP does exist, it is undoubtedly a gross oversimplification to use a simple non-null binomial model for these experiments. In addition to potential differences in ability among subjects, there were also observed differences due to dynamic versus static targets, whether or not the sender was a friend, and how the receiver scored on measures of extraversion. All of these differences were anticipated in advance and could be incorporated into models as covariates.

It is nonetheless instructive to examine the Bayes factor computed by Bayarri and Berger for the simple non-null binomial model. First, the observed anomalous effects would be less interesting if the Bayes factor was small for reasonable values of r, as it was for the random number generator experiments analyzed by Jefferys (1990), most of which purported to measure psychokinesis instead of ESP. Second, the Bayes factor provides a rough measure of the strength of the evidence against the null hypothesis and is a much more sensible summary than the p-value. The Bayes factors provided by Bayarri and Berger are probably more conservative, in the sense of favoring the null hypothesis, than those that would result from priors elicited from parapsychologists, but are probably reasonable for those who know nothing about past observed effects. I expect tht most parapsychologists would not opt for a prior symmetric around chance, but would still choose one with some mass below chance. The final reason it is instructive to examine these Bayes factors is that they provide a quantitative challenge to skeptics to be explicit about their prior probabilities for the null and alternative hypotheses.

Dawson discussed the use of more complex Bayesian models for the analysis of the autoganzfeld data. She proposed a hierarchical model where the number of successes for each experiment followed a binomial distribution with hit rate p_i, and logit(p_i) came from a normal distribution with noninformative priors for the mean and variance. She then expanded this model to include heavier tails by allowing an additional scale parameter for each experiment. Her rationale for this expanded model was that there were clear outlier series in the data.

The hierarchical model proposed by Dawson is a reasonable place to start given only that there were several experiments trying to measure the same effect, conducted by different investigators. In the autoganzfeld database, the model could be expanded to incorporate the additional information available. Each experiment contained some sessions with static targets and some with dynamic targets, some sessions in which the sender and receiver were friends and others in which they were not and some information about the extraversion score of the receiver. All of this information could be included by defining the individual session as the unit of analysis, and including a vector of covariates for each session. It would then make sense to construct a logistic regression model with a component for each experiment, following the model proposed by Dawson, and a term $X\beta$ to include the covariates. A prior distribution for β could include information from earlier ganzfeld studies. The advantage of using a Bayesian approach over a simple logistic regression is that information could be continually updated. Some of the recent work in Bayesian design could then be incorporated so that future trials make use of the best conditions.

Several of the discussants addressed the concept of replication. I agree with Mosteller's implication that it was unwise for the audience in my seminar to respond to my replication questions so quickly, and that was precisely my point. Most nonstatisticians do not seem to understand the complexity of the replication question. Parenthetically, when I posed the same scenario to an audience of statisticians, very few were willing to offer a quick opinion.

Bayarri and Berger provided an insightful discussion of the purpose of replication, offering quantitative answers to questions that were implicit in my discussion. Their analyses suggest some alternatives to power analysis that might be considered when designing a new study to try to replicate a questionable result.

Morris addressed the question of what constitutes a replication of a meta-analysis. He distinguished between exact and conceptual replications. Using his distinction, the autoganzfeld meta-analysis could be viewed as a conceptual replication of the earlier ganzfeld meta-analysis. He noted that when such a conceptual replication offers results similar to those of the original meta-analysis, it lends legitimacy to the original results, as was the case with the autoganzfeld meta-analysis.

Greenhouse and Morris both noted the value of meta-analysis as a method of comparing different conditions, and I endorse that view. Conditions found to produce different effects in one meta-analysis could be explicitly studied in a conceptual replication. One of the intriguing results of the autoganzfeld experiments was that they supported the distinction between effect sizes for dynamic versus static targets found in the earlier ganzfeld work, and they supported the relationship between ESP and extraversion found in the meta-analysis by Honorton, Ferrari and Bem (1990).

Most modern parapsychologists, as indicated by Morris, recognize that demonstrating the validity of their preliminary findings will depend on identifying and utilizing "moderator variables" in future studies. The use of such variables will require more complicated statistical models than the simple binomial models used in the past. Further, models are needed for combining results from several different experiments, that don't oversimplify at the expense of lost information.

In conclusion, the anomalous effect that persists throughout the work reviewed in my paper will be better understood only after further experimentation that takes into account the complexity of the system. More realistic, and thus more complex, models will be needed to analyze the results of those experiments. This presents a challenge that I hope will be welcomed by the statistics community.

ADDITIONAL REFERENCES

ALLISON, P. (1979). Experimental parapsychology as a rejected science. *The Sociological Review Monograph* **27** 271–291.

BARBER, B. (1961). Resistance by scientists to scientific discovery. *Science* **134** 596–602.

BERGER, J. O. and DELAMPADY, M. (1987). Testing precise hypotheses (with discussion). *Statist. Sci.* **2** 317–352.

CHUNG, F. R. K., DIACONIS, P., GRAHAM, R. L. and MALLOWS, C. L. (1981). On the permanents of compliments of the direct sum of identity matrices. *Adv. Appl. Math.* **2** 121–137.

COCHRAN, W. G. (1954). The combination of estimates from different experiments. *Biometrics* 10 101-129.

COLLINS, H. and PINCH, T. (1979). The construction of the paranormal: Nothing unscientific is happening. *The Sociological Review Monograph* 27 237-270.

COLLINS, H. M. and PINCH, T. J. (1982). *Frames of Meaning: The Social Construction of Extraordinary Science*. Routledge & Kegan Paul, London.

CORNFIELD, J. (1959). Principles of research. *American Journal of Mental Deficiency* 64 240-252.

DEMPSTER, A. P., SELWYN, M. R. and WEEKS, B. J. (1983). Combining historical and randomized controls for assessing trends in proportions. *J. Amer. Statist. Assoc.* 78 221-227.

DIACONIS, P. and GRAHAM, R. L. (1981). The analysis of sequential experiments with feedback to subjects. *Ann. Statist.* 9 236-244.

FISHER, R. A. (1932). *Statistical Methods for Research Workers*, 4th ed. Oliver and Boyd, London.

FISHER, R. A. (1935). Has Mendel's work been rediscovered? *Ann. of Sci.* 1 116-137.

GALTON, F. (1901-2). Biometry. *Biometrika* 1 7-10.

GREENHOUSE, J., FROMM, D., IYENGAR, S., DEW, M. A., HOLLAND, A. and KASS, R. (1990). Case study: The effects of rehabilitation therapy for aphasia. In *The Future of Meta-Analysis* (K. W. Wachter and M. L. Straf, eds.) 31-32. Russell Sage Foundation, New York.

GRIFFIN, D. (1988). Intuitive judgment and the evaluation of evidence. In *Enhancing Human Performance: Issues, Theories and Techniques Background Papers—Part I*. National Academy Press, Washington, D.C.

HANSEN, G. (1990). Deception by subjects in psi research. *Journal of the American Society for Psychical Research* 84 25-80.

HUNTER, J. and SCHMIDT, F. (1990). *Methods of Meta-Analysis*. Sage, London.

IYENGAR, S. and GREENHOUSE, J. (1988). Selection models and the file drawer problem (with discussion). *Statist. Sci.* 3 109-135.

LOUIS, T. A. (1984). Estimating an ensemble of parameters using Bayes and empirical Bayes methods. *J. Amer. Statist. Assoc.* 79 393-398.

MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 22 719-748.

MORRIS, C. (1983). Parametric empirical Bayes inference: Theory and applications (rejoinder) *J. Amer. Statist. Assoc.* 78 47-65.

MORRIS, R. L. (1986). What psi is not: The necessity for experiments. In *Foundations of Parapsychology* (H. L. Edge, R. L. Morris, J. H. Rush and J. Palmer, eds.) 70-110. Routledge & Kegan Paul, London.

MOSTELLER, F. and BUSH R. R. (1954). Selected quantitative techniques. In *Handbook of Social Psychology* (G. Lindzey, ed.) 1 289-334. Addison-Wesley, Cambridge, Mass.

MOSTELLER, F. and CHALMERS, T. (1991). Progress and problems in meta-analysis. *Statist. Sci.* To appear.

OTERI, L., ed. (1975). *Quantum Physics and Parapsychology*. Parapsychology Foundation, New York.

PINCH, T. J. and COLLINS, H. M. (1984). Private science and public knowledge: The Committee for the Scientific Investigation of Claims of the Paranormal and its use of the literature. *Social Studies of Science* 14 521-546.

PLATT, J. R. (1964). Strong inference. *Science* 146 347-353.

ROSENTHAL, R. (1966). *Experimenter Effects in Behavioral Research*. Appleton-Century-Crofts, New York.

ROSENTHAL, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin* 86 638-641.

RYAN, L. M. and DEMPSTER, A. P. (1984). Weighted normal plots. Technical Report 394Z, Dana-Farber Cancer Inst., Boston, Mass.

SAMANIEGO, F. J. and UTTS, J. (1983). Evaluating performance in continuous experiments with feedback to subjects. *Psychometrika* 48 195-209.

SMITH, M. and GLASS, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist* 32 752-760.

WACHTER, K. (1988). Disturbed by meta-analysis? *Science* 241 1407-1408.

WEST, M. (1985). Generalized linear models: Scale parameters, outlier accommodation and prior distributions. In *Bayesian Statistics 2* (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.) 531-558. North-Holland Amsterdam.

APPENDIX H
AN ASSESSMENT OF THE ENHANCED HUMAN PERFORMANCE PROJECT

 SECRET

copy to 189 6/28

THE ENHANCED HUMAN PERFORMANCE PROJECT: AN ASSESSMENT OF THE EFFORT TO DATE

PROJECT REVIEW GROUP
14 APRIL, 1987

At the request of MG Philip K. Russell, MC, Commander, United States Army Medical Research and Development Command, the following individuals met at the Pentagon on 6 March 1987 to assess the work of the Enhanced Human Performance Project:

Ms. Amoretta Hoeber, TRW
Dr. Jack Vorona, DIA
Dr. Michael A. Wartell, Humboldt State University
Dr. Nick Yaru, Consultant (Chairman)
Dr. Chris Zarafonetis, Biomedical R&D, Inc.

Others in attendance at this meeting included:

BG Richard T. Travis, MC, Deputy Commander, USAMRDC
Col. Philip Sobocinski, MSC, Special Assistant for Biotechnology
Col. Peter J. McNelis, MSC, Project Manager/COR
Mrs. Jean Smith, Principal Assistant Responsible for Contracting
Dr. Edwin C. May, SRI, Principal Investigator

In preparation for this meeting, copies of all Project reports for Fiscal Year 1986 along with the Scientific Oversight Committee's comments regarding these reports and the contractor's responses to the comments were forwarded to each of the above-mentioned individuals for their review.

The Project Review Group was asked, via correspondence (MG Russell, 12 January 1987; Col. McNelis, 12 February 1987) and by BG Travis in his welcoming remarks at the meeting, to address the following questions concerning the Project:

1. Is the science underlying this research effort essentially sound?

2. Does the evidence to date support the existence of an anomaly?

3. What is the potential value of this effort to the DOD?

SECRET

4. Is the research focus and level of effort appropriate?

The agenda for the meeting is attached as Enclosure 1. Following a presentation of the Project's historical antecedents, the questions listed above provided the structure for a discussion of: FY 1986 research tasks and results, the overall plan underlying the FY 1986, effort and possible modifications of the plan for follow-on work.

The Review Group's responses to the preceding questions and their recommendations for the Project will be presented in turn. It should be noted that there was unanimity among the members of the Review Group with regard to these responses.

1. Is the science sound?

   The individual experiments conducted during Fiscal Year 1986 appear to be scientifically sound. The primary contractor's response to comments of the Scientific Oversight Committee (SOC) leads this Review Group to conclude that the scientific quality of the effort is under continual qualified scrutiny, and immediate adjustments are made by the researchers to insure that that quality continues. Additionally, appropriate community-wide symposia such as the Theory and Proof of Principle conferences projected for FY 1987 will enhance that quality.

2. Is there an anomaly?

   The results of experiments conducted by this Project during FY 1986, as well as other reports of previous operational related research, lead this Review Group to conclude that a natural anomaly exists, which we will refer to as Remote Viewing.

3. Is it worthwhile?

   The Review Group believes that progress is being made in understanding this anomaly and that continuation of the effort is not only warranted, but entirely appropriate and strongly recommended.

   Should Remote Viewing be predictably reproducible and its mechanisms, parameters and physiological correlates understood, there would be a number of significant applications for the DoD. Current user agencies have reported utilizing the present technology with positive results.

4. Is the direction and emphasis appropriate?

The Review Group believes that the probability of success in demonstrating and explaining a phenomenon known as Remote Action is less than the probability of success for the Remote Viewing phenomenon. Rather than continuing to explore both phenomena at equal levels of effort, it is recommended that the results of this year's (FY87) effort be critically reviewed and those areas that demonstrate the most promise be exploited and those that do not be terminated. The focus then would be less diffuse and more vertical as the more productive pathways are emphasized.

This should not be considered an economy measure, however, since the vertical effort should be assured of adequate resources to accomplish its more definitive tasks.

The Review Group also recommends that the Project should clarify its use of the terms: global/conceptual replication (i.e., other labs evidence the phenomena without following the same protocol), exact/technical replication (i.e., phenomena evidenced in other labs following the same protocol with other subjects and other targets), and reproducibility (i.e., phenomena evidenced by the same subjects over time utilizing the same randomly ordered target set). With this in mind, it is recommended that an effort be made to enhance the reproducibility of the phenomena by identifying and utilizing especially talented individuals. It is believed that this pool of talented subjects would also aid in isolating neurophysiological correlates and mechanisms.

It is also recommended that one or two other secure labs be identified to carry out exact/technical replication of the most promising experiments conducted by the primary contractor.

Overall, the current breadth of experiments selected to demonstrate and explicate the phenomena is appropriate, as is the present level of effort assigned to each of these experiments.

In summary, the Project Review Group has determined to its satisfaction that the work of the Enhanced Human Performance Project is scientifically sound, appropriately managed and monitored, and is providing valuable insight into the nature of an anomaly which could have a significant impact on the DoD.

Dr. Nick Yaru

Dr. Nick Yaru, Chairman
Project Review Group

APPENDIX I
IN-HOUSE STAFFING REQUIREMENTS

(S/NF/SG/LIMDIS) An analysis of the PAG-TA functions necessary to support the achievement of the long-range goals indicate four major functional areas which must be supported. Within each functional area, personnel requirements can be identified. A complicating factor, however, is the fact that some of the functional areas (such as remote viewing (RV), Intelligence Analysis, and ADP support) are highly specialized and require full-time dedicated personnel.

1. (S/NF/SG/LIMDIS) RV Activities: RV activities can be grouped into the following major areas:

a. Participate in R & D activities with the external R&D contractor
b. Viewer Training (both in-house and with the external R&D contractor)
c. Operational Activities

(S/NF/SG/LIMDIS) It is difficult to project personnel requirements for this functional area, primarily because the projected level of operational activity is currently unknown. Based on the past level of operational tasking, it is anticipated that up to six personnel could be required. Five of the people would be involved in operational activities as well as participating in support of the R&D activities to be conducted by the external Contractor. One additional person would be designated to participate in operational and research support activities on a part-time basis but would devote most of his time to developing a training program and conducting training of new personnel and identification/selection of potential viewers. Due to the specialized nature of RV, this person needs to be a qualified viewer and not merely an administrative person. It should also be kept in mind that it takes approximately one year to train a viewer to operational status.

2. (U) Foreign Intelligence Assessment: Support of this functional area may be grouped into the following activities:

a. Data source identification/collection
b. Construction of Foreign Activities Data Base

c. Analysis
d. Production of finished intelligence assessments

(U) To adequately meet the requirements of this functional area, two full-time personnel will be required: a Senior Intelligence Officer (SIO) and an Intelligence Technician (IT). In order to maintain strict protocol requirements, these personnel should not function as operational viewers.

(U) The IT would identify potential sources of data, collect the data, support the construction of the Intelligence database and input the required data,and assist in the preparation of intelligence assessments. The SIO should be an all-source Scientific and Technical Intelligence analyst and would be responsible for the identification of collection requirements, the analysis of intelligence data, and the production of finished intelligence assessments on a world-wide basis.

3. (S/NF) ADP Support: Over the period of time covered by this Plan, the ADP support activities of PAG-TA are anticipated to rise dramatically, requiring one full-time person to function as an ADP system administrator. Several factors justify this position:

a. (S/NF) PAG-TA is currently in the process of upgrading its ADP system to include the acquisition of a Unix-based SUN workstation which will not only serve as the main system element, but will also be used to construct the Intelligence and the R&D databases, serve as the communications link to the external Contractor, and support the operation of special PAG-TA research equipment. Specific areas requiring specialized technical attention include:

(1) Operating system(s)
(2) Potential LAN(s) administration
(3) Database construction/maintenance
(4) Language compiler(s)
(5) Peripherals
(6) Equipment interfaces
(7) Data communications
(8) System modifications/upgrades
(9) Development of special purpose software to support the PAG-TA mission

b. (C) PAG-TA is located some distance from the main Agency computer support facilities. Should the PAG-TA system experience problems or failures, the system would be down

until someone from the main facility could travel to the PAG-TA location to effect repairs, resulting in a loss of productivity during the wait period. Also, any system modification/upgrades

would have to depend on the schedule of qualified personnel, again resulting in loss of productivity. Therefore; it is essential that a person will the necessary computer science skills be physically located at the PAG-TA facility.

4. (S/NF/SG/LIMDIS) <u>Branch Administration</u>: Tasks in this functional area may be grouped as follows:

a. Word Processing
   (1) Electronic Filing
   (2) Management Support
   (3) Security Administration
   (4) Report Generation/Document Preparation
   (5) RV Tasking
   (6) Generation of RV Target Pools

b. Project/Contract Management

c. Collection Management

d. Ft. Meade Interface/Facilities

5. (S/NF/SG/LIMDIS) Tasks in this area will require three to four personnel--a Branch Chief, a person functioning as an Assistant Branch Chief (probably the SIO), a Secretary and, possibly, a Collection Manager (unless this can be done on an "as required" basis by other Branch personnel). The Branch Chief and SIO should have experience in project/contract management, primarily to deal with external research/support contracts, as well as the ability to interface with the academic community and professional organizations engaged in parapsychological activities in addition to overall management skills associated with managing a Branch-size organization.

(C) Based on this evaluation, a total of 11-12 personnel could be required to effectively achieve PAG-TA goals. No attempt has been made to identify the personnel as either military or civilian. This represents an increase of 1-2 personnel over the current authorization. However; it may be more desirable to keep the manning level at current strength (10 authorized/7 assigned) and adjust the existing skill mix at PAG-TA to more effectively meet anticipated programmatic demands through personnel transfers/reassignments.