(Original publication and copyright: *Journal of Parapsychology*, 2004, Volume 68, pages 157-167)

A PROPOSAL AND CHALLENGE FOR PROPONENTS AND SKEPTICS OF PSI

By J.E. Kennedy

ABSTRACT: Pharmaceutical research provides a useful model for doing convincing research in situations with intense, critical scrutiny of studies. The protocol for a "pivotal" study that is used for decision-making is reviewed by the FDA before the study is begun. The protocol is expected to include a power analysis demonstrating that the study has at least a .8 probability of obtaining significant results with the anticipated effect size, and to specify the statistical analysis that will determine the success of the experiment, including correction for multiple analyses. FDA inspectors often perform audits of the sites where data are collected and/or processed to verify the raw data and experimental procedures. If parapsychological experiments are to provide convincing evidence, power analyses should be done at the planning stage. A committee of experienced parapsychologists, moderate skeptics, and a statistician could review and comment on protocols for proposed "pivotal" studies in an effort to address methodological issues before rather than after the data are collected. The evidence that increasing sample size does not increase the probability of significant results in psi research may prevent the application of these methods and raises questions about the experimental approach for psi research.

In recently reading the 1988 Office of Technology Assessment report on experimental parapsychology (Office of Technology Assessment, 1989), I was struck by two topics: the optimism for meta-analyses and the suggestion that proponents of psi and skeptics should form a committee to evaluate and guide research.

In the decade and a half since this report, the use of meta-analyses has become more common, and the controversial aspects and limitations have become more clear. Meta-analysis is ultimately post hoc data analyses when researchers have substantial knowledge of the data. Evaluation of the methodological quality of a study is done after the results are known, which gives opportunity for biases to affect the meta-analysis. Different strategies, methods, and criteria can be utilized, which can give different outcomes and opportunity for selecting outcomes consistent with the analyst's expectations. The meta-analysis results can vary as new studies become available, which raises the possibility of optional stopping and selective reporting. The various controversies over meta-analyses with the ganzfeld demonstrate these issues (Milton, 1999; Schmeidler & Edge, 1999; Storm, 2000).

Bailar (1997) described similar conclusions from the experience with meta-analysis in medical research:

> It is not uncommon to find that two or more meta-analyses done at about the same time by investigators with the same access to the literature reach incompatible or even contradictory conclusions. Such disagreement argues powerfully against any notion that meta-analysis offers an assured way to distill the "truth" from a collection of research reports. (p. 560)

The research strategies and procedures in parapsychology stand in marked contrast with pharmaceutical research, through which I now earn my livelihood. The level of planning, scrutiny, and resulting evidence is much higher in pharmaceutical research than in most academic research, including parapsychology.

Pharmaceutical research offers a useful model for providing convincing experimental results in controversial situations. Key aspects of this research process are described below.

Basic Pharmaceutical Research

A company that wants to provide convincing evidence that a new product is effective begins by doing a few or several small exploratory or pilot studies. These are called Phase 1 and Phase 2 studies and are used to develop the methods of administering the product and effective dose as well as providing initial evidence for the benefits and potential adverse effects in humans.

When the researchers believe that they know the effective dose and can deliver it reliably, and that the effectiveness may be sufficient to be profitable, they plan a "pivotal Phase 3" study. This is a study that is intended to provide convincing evidence and is normally a randomized experiment. The study protocol describes the study procedures, specific data items to be collected, patient population, sample size, randomization, and planned analyses. The general statistical methods expected by the U.S. Food and Drug Administration (FDA) and corresponding agencies in many other countries are described in "Guidance for Industry: E9 Statistical Principles for Clinical Trials" (available for downloading at no charge from [the FDA updated link updated after publication]). This document is

excellent guidance for anyone doing experimental research in controversial settings and is part of the international standards for pharmaceutical research that are being developed by the International Conference on Harmonisation (ICH).

The protocol is expected to include a power analysis demonstrating that the study sample size has at least .8 to .9 probability of obtaining significant results if the effects are of the assumed magnitude. Sensitivity analyses exploring a variety of deviations from the assumptions in the power analyses are recommended, and are important for the company as well as for the FDA.

The single "primary variable" that will determine the success of the study is specified, as is the specific statistical analysis, including any covariates. If there is more than one primary outcome analysis, then correction for multiple analyses is expected to be specified in the protocol. There are usually several "secondary variables" that are used as supporting evidence and are handled more leniently than the primary outcome, but still all variables and the basic analysis plan should be specified in the protocol.

Prior to beginning the study, the protocol is submitted to the FDA for review and comments. This normally involves discussions and revisions. The company is not legally required to follow the FDA's suggestions at this stage, but it is clearly wise to reach agreement before starting the study.

For most products, two pivotal Phase 3 studies are required. Both follow the criteria and process described above. The two studies may be done sequentially or concurrently. If the results do not turn out as expected, additional studies may be needed.

When the studies are completed and the company is ready to submit the application for approval, the full study reports for all studies (including Phases 1 and 2) are submitted to the FDA along with listings of all data and usually electronic copies of the data. There is also a section on "integrated analyses" that combines the data from the studies. The FDA increasingly evaluates applications by performing its own analyses of the electronic data.

It is common for the FDA to send inspectors to the site(s) where data were collected and/or processed to verify the raw data and review the procedures for data collection and processing. This site audit specifically verifies that the procedures stated in the protocol were followed and that the raw data records match the computer database to a high degree of accuracy. If there are discrepancies or if the data collection involved particular reliance on electronic data capture, the

audit may include evaluating the data processing systems and programs. Usually, security and restricted access to the data and relevant data processing systems are also significant issues for site audits. Companies usually have internal quality control procedures that double and triple check all research activities in anticipation of being audited.

After the FDA has all relevant information, it may make a decision internally, or it may convene a scientific advisory board that reviews the information, asks questions to the company and the FDA, and makes recommendations. An advisory board is likely if the studies produce results that are equivocal. Any deviation from the protocols in procedure or analysis must be explained and can be a significant obstacle.

It may be worth noting that workers in pharmaceutical research generally do not take offense that everything they do, including the simplest tasks, is questioned and double or triple checked. In fact, these quality control efforts reveal a surprising number of mistakes and oversights. The attitude quickly becomes one of working together as a team to overcome the human tendency to make mistakes. The redundant checking is taken as an indication of how important the project is. Pharmaceutical researchers generally view academic research as having much lower quality and find that it takes substantial effort to retrain academic researchers to meet the higher standards.

A Proposal for Parapsychology

Given that the research processes described above are my standard of reference now, I do not expect that the current meta-analysis approaches in parapsychology will provide convincing evidence for even mild skeptics. The meta-analysis strategy in parapsychology seems to be to take a group of studies in which 70% to 80% or more of the studies are not significant and combine them to try to provide evidence for an effect. This is intrinsically a post hoc approach with many options for selecting data and outcomes.

More generally, the usual standards of academic research may not be optimal for addressing controversial, subtle phenomena such as psi. Because of the relatively high noise levels in academic research, widespread independent replication is usually required for evidence to become convincing. Phenomena that are more subtle and difficult

to replicate may require a lower noise level for convincing evidence and scientific progress.

It appears to me that preplanned analysis of studies with sufficient sample size to reliably obtain significant results is necessary to provide convincing experimental results and meaningful probability values in controversial settings such as parapsychology. Sample sizes should be set so that the probability of obtaining significant results is at least .8 given a realistic psi effect. This is a substantial change from the current practice in which studies are done with little regard for statistical power and only about 20% to 30% are significant, which results in controversy and speculation about whether the predominately negative results are due to a lack of psi or a lack of sample size. Performing a prospective power analysis is simply doing what statisticians have long recommended.

If the claims that meta-analyses results provide evidence for psi are actually valid, then this approach of prospective power analysis and study planning will be successful. If this approach will not work, then the application of statistical methods in parapsychology, including meta-analyses, will not be convincing.

From my perspective now, it would make good sense to form a committee consisting of experienced parapsychologists, moderate skeptics, and at least one statistician to review and comment on protocols for pivotal experiments prior to the experiments being carried out. The committee could also do independent analyses of data, verify that the analyses comply with those planned in the protocol, and perhaps sometimes do site inspections. The possibility of a detailed, on-site, critical audit of the experimental procedure and results provides a healthy perspective on methodology. It would be valuable to have this option available even if it is rarely or never used.

The idea of a registry to distinguish between exploratory and confirmatory experiments has been suggested several times over the years (e.g., Hyman & Honorton, 1986; also see comments in Schmeidler & Edge, 1999). This strategy would allow researchers more freedom to do exploratory studies as long as the confirmatory or pivotal studies are formally defined in advance.

The present proposal is an extension of the registry idea that would also attempt to resolve much of the methodological controversy before rather than after a study is carried out. The most efficient strategy to obtain a consensus is to have those people who are critical provide input and agree on the research plan from the beginning. The net

effort to carry out a study and answer criticisms may actually be less, and the final quality of evidence should be substantially higher.

This strategy is consistent with the idea that only certain experimenters can be expected to obtain positive results and does not require that any experimenter, no matter how skeptical, must be able to consistently obtain significant results. Thus, this strategy is reasonably consistent with the known characteristics of psi research.

This strategy also allows starting with a clean slate for evaluating research. The studies that comply with this process can stand as a separate category to determine whether there is evidence for psi. Given the higher quality of each pivotal study, there would be less need for many replications, and experimenters would have more freedom to capitalize on the novelty effect of starting new studies. A pre-specified analysis and criteria could be set for determining whether a group of studies provides overall evidence for psi. This could focus on certain experimenters with a track record of success, rather than expecting any and all experimenters to be successful.

Challenges for Skeptics

I expect that many of the more extreme skeptics will be hesitant to participate, or more likely, simply never be able to agree prospectively that a protocol is adequate. These skeptics appear happy to devote many hours to after-the-fact speculations and listing deficiencies in past experiments, and they claim that convincing experiments are certainly possible, but they will find it very uncomfortable to specify prospectively that a study design is adequate to provide evidence for psi. These skeptics must recognize that their beliefs, arguments, and behavior are not scientific.

The members of the committee would have to be people who agree with the principle that experimental research methods can be used to obtain meaningful evidence (pro or con) in parapsychological research, and they would have to be willing to support and adhere to the standards of scientific research, no matter what the outcome.

These proposals would also limit the skeptical practice of doing a large number of post hoc internal analyses for studies that are significant and then presenting selected results as worrisome. If a skeptic (or proponent) believes that certain internal consistencies are important, then appropriate analyses, including adjustment for multiple analyses, can be pre-specified in the protocol. Post hoc data

scrounging by skeptics would be recognized as a biased exercise with minimal value, as is post hoc scrounging of nonsignificant data to try to find supportive results.

Challenges for Proponents

Parapsychologists may be skeptical of these proposals because they believe that psi is not sufficiently reliable to carry out this type of research program. Attempts to apply power analyses to a phenomenon that has the experimenter differences and declines found in psi research bring these reliability issues into focus (Kennedy, 2003a). However, if these reliability issues preclude useful experimental planning, then the effects are not sufficiently consistent for convincing scientific conclusions.

The declines in effects across experiments are particularly problematic for planning studies and prospective power analysis. For example, the first three experiments on direct mental interactions with living systems carried out by Braud and Schlitz each obtained consistent, significant effects with 10 sessions (reviewed in Braud & Schlitz, 1991). Six of the subsequent experiments had 32 to 40 sessions, which prospectively would be expected to have a high probability of success given the effects in the first three studies. However, only one of the six experiments reached statistical significance.

The declining effects and corresponding need for large sample sizes makes research unwieldy, expensive, and prone to internal declines. For the 33% hit rate found in the early ganzfeld research (Utts, 1986), a sample size of 192 is needed to have a .8 probability of obtaining a .05 result one-tailed.[1] Broughton and Alexander (1997) carried out a ganzfeld experiment with a preplanned sample size of 150 trials. The overall results were nonsignficant and there was a significant internal decline. Similarly, Wezelman and Bierman (1997) reported overall nonsignificant results and significant declines over 236 ganzfeld trials obtained from 6 experiments at Amsterdam. On the other hand, Parker (2000) reported overall significant results and no declines in 150 ganzfeld trials obtained from 5 experiments. Likewise, Bem and Honorton (1994) reported overall significant results without declines in 329 trials obtained from 11 experiments. These latter three reports apparently summarized an accumulation of studies that were carried out without pre-specifying the combined

samples size, but they do raise the possibility that preplanned studies with adequate sample size may be possible without internal decline effects.

However, the overriding dilemma for power analysis in psi research is that increasing the sample size apparently does not increase the probability of obtaining significant results. In addition to the Braud and Schlitz (1991) studies described above, the *z* score or significance level was unrelated to sample size in meta-analyses of RNG studies (Radin & Nelson, 2000) and early ganzfeld studies (Honorton, 1983). Equivalently, effect size was inversely related to sample size in RNG studies (Steinkamp, Boller & Bosch, 2002) and later ganzfeld studies (Bem & Honorton, 1994).[2] Contrary to the basic assumptions for statistical research, sample size does not appear to be a significant factor for the outcome of psi experiments.

In medical research, finding larger average effect sizes in studies with smaller sample sizes is an established symptom of methodological bias, such as publication bias, selection bias, methodological quality varying with study size, and selected subject populations (Egger, Smith, Schneider, & Minder, 1997). These biases are evaluated by examining a plot of effect size against sample size to see if the distribution differs from the expected symmetric inverted funnel shape. This method has revealed apparent biases in meta-analyses for which the conclusions were contradicted by subsequent large studies (Egger, Smith, Schneider, & Minder, 1997).

However, for psi experiments, the inverse relationship between effect size and sample size may be a manifestation of goal-oriented psi experimenter effects (Kennedy, 1994; 1995) and decline effects (Kennedy, 2003b) rather than methodological bias. Whether or not it is caused by psi, this relationship and the associated failure to find larger *z* scores with larger sample sizes have ominous implications for planning and interpreting psi experiments.

For these and other reasons,[3] I personally doubt that psi has the properties needed for experimental research as historically attempted in parapsychology and as assumed for the proposals here. Alternative research strategies and more innovative statistical methods may be needed (Kennedy, 1994, 1995, 2003b). The hypothesis of goal-oriented psi experimenter effects appears to be a good fit to the data. Psi effects by definition do not conform to known physical laws, but the psi effects in experiments are assumed to conform to the standard physical properties and processes of experimental research. It does not surprise me to find evidence that this assumption is incorrect.

However, I recognize that the majority of parapsychologists, particularly proponents of meta-analysis, probably dispute my views on this. The proposals presented here appear to me to be the optimum way to directly evaluate whether the assumptions of experiments and meta-analyses are valid for parapsychology, and to advance the field if these methods apply.

Whether one believes that the standard statistical methods are adequate for psi research or that more innovative approaches need to be developed, the level of planning in research programs must be substantially increased if parapsychology is to advance beyond its present state of uncertainty, controversy, and reliance on post hoc analyses.

References

Bailar, J.C. (1997). The promise and problems of meta-analysis. *New England Journal of Medicine*, **337**, 559-561.

Bem, D.J., & Honorton, C. (1994). Does psi exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin*, **115**, 4-18.

Braud, W.G., & Schlitz, M.J. (1991). Consciousness interactions with remote biological systems: Anomalous intentionality effects. *Subtle Energies*, **2** 1-46.

Broughton, R.S., & Alexander, C.H. (1997). Autoganzfeld II: An attempted replication of the PRL ganzfeld research. *Journal of Parapsychology*, **61**, 209-226.

Brown, R.H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine*, **14**, 1933-1940.

Egger, M., Smith, G.D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple graphical test. *British Medical Journal*, **315**, 629-634.

Honorton, C. (1983). Response to Hymans's critique of ganzfeld studies. In W.G. Roll, J. Beloff & R.A. White (Eds.), *Research in parapsychology 1982* (pp. 23-26). Metuchen, NJ: Scarecrow Press.

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

Lachin, J.M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, **2**, 93-113.

Kennedy, J.E. (1994). Exploring the limits of science and beyond: Research strategy and status. *Journal of Parapsychology*, **58**, 59-77.

Kennedy, J.E. (1995). Methods for investigating goal-oriented psi. *Journal of Parapsychology*, **59**, 47-62.

Kennedy, J.E. (2003a). Letter to the editor. *Journal of Parapsychology*, **67**, 406-408.

Kennedy, J.E. (2003b). The capricious, actively evasive, unsustainable nature of psi: A summary and hypotheses. *Journal of Parapsychology*, **67**, 53-74.

Milton, J. (1999). Should ganzfeld research continue to be crucial in the search for a replicable psi effect? Part I. Discussion paper and introduction to an electronic-mail discussion. *Journal of Parapsychology*, **63**, 309-333.

Office of Technology Assessment (1989). Report of a workshop on experimental parapsychology. *Journal of the American Society for Psychical Research*, **83**, 317-339.

Parker, A. (2000). A review of the ganzfeld work at Gothenburg University. *Journal of the Society for Psychical Research*, **64**, 1-15.

Radin, D., & Nelson, R. (2000). Meta-analysis of mind-matter interaction experiments: 1959 to 2000. Unpublished manuscript, Boundary Institute, Los Altos, CA, and Princeton Engineering Anomalies Research, Princeton University.

Schmeidler, G.R., & Edge, H. (1999). Should ganzfeld research continue to be crucial in the search for a replicable psi effect? Part II. Edited ganzfeld debate. *Journal of Parapsychology*, **63**, 335-388.

Steinkamp, F., Boller, E., & Bosch, H. (2002). Experiments examining the possibility of human intention interactions with random number generators: A preliminary meta-analysis [Abstract]. *Journal of Parapsychology*, **66**, 238-239.

Storm, L. (2000). Research note: Replicable evidence of psi: A revision of Milton's (1999) meta-analysis of ganzfeld databases. *Journal of Parapsychology*, **64**, 411-416.

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

Wezelman, R., & Bierman, D.J. (1997). Process oriented ganzfeld research in Amsterdam: Series IVb, emotionality of target material, series V and series VI: Judging procedure and altered states of

consciousness. *Proceedings of Presented Papers: The Parapsychological Association 40th Annual Convention,* 477-491.

Notes

1. Internet sites with interactive power calculations are available (e.g., http://calculators.stat.ucla.edu/powercalc/). Lachin (1981) gives relatively simple equations for power calculations that can be done with a calculator. Brown (1995) provides useful information on using data from pilot studies to estimate variances for power calculations.

2. A basic principle of probability theory is that the *z* score is expected to increase with the square root of sample size. This is the basis for power analysis and for using the *z* score divided by the square root of sample size as an effect size measure that is expected to be unrelated to sample size. If the *z* score does not increase with sample size, then effect size will be inversely related to sample size. Among other implications, this anomalous result defeats the meaning and purpose of effect size measures in meta-analysis.

3. The consistent evidence that majority-vote studies in psi research do not have the properties assumed for standard signal enhancement methods and appear to be dominated by the experimenter's wants and expectations (Kennedy, 1995) provides strong evidence for this point. Power analysis and majority-vote methods derive from the same basic principles of probability theory.

Return to: Paranormal Phenomena Articles