# HOW TO EVALUATE AND IMPROVE THE REPLICABILITY OF PARAPSYCHOLOGICAL EFFECTS

CHARLES HONORTON

*Replicability: A Ubiquitous Problem*

Any realistic appraisal of replicability in parapsychology must begin with the recognition that replicability problems are not unique to psi research, but exist in varying degrees in many different areas of research in the biological, behavioral and social sciences.

Not even the "hard sciences" are immune to replicability problems, as was illustrated by British sociologist of science Harry Collins in his study of the "Transversely Excited Atmospheric Pressure CO2 Laser," or TEA Laser (Collins, 1974). This device was invented in Canada in 1968 and Collins located seven British laboratories who had built or were trying—with varying degrees of success—to build TEA lasers and he interviewed most of the British scientists who were directly involved. He concluded: "The transfer of knowledge appeared to be a capricious process which nearly always involved a period of face-to-face contact between R(eplicating) S(cientist) and a scientist who had already built a (working) laser. Thus though the knowledge could travel along a chain of intermediaries from O(riginating) S(cientist) to RS, these intermediaries had to be replicators—culturally competent—themselves, not just the passive carriers of algorithmical type information. Where scientists tried to build a laser on written information, or information provided by third parties who were not themselves replicators, they failed. Furthermore, even prolonged personal contact was not necessarily sufficient. Some scientists could not succeed in building a TEA laser and eventually abandoned the project in spite of their good access to sources of help" (Collins, 1978, p. 9).

Social and behavioral science commentators have expressed growing concern over the replicability of findings in a number of research areas. This concern is an outgrowth of a variety of factors ranging from professional publication practices to the intrinsic variability of

human behavior. Surveys of publication practices in American psychological journals have shown that while 94–97 percent of knowledge claims in psychology are made on the basis of statistical significance tests, less than 1 percent of the studies represent replications of earlier findings (Sterling, 1959; Bozarth and Roberts, 1972). A recent article in *American Psychologist* by Sommers and Sommers (1983) states the situation in psychology succinctly: "A major structural problem in psychological science is the lack of any requirement for replication prior to publication. Some journals will not accept replications. As a result, many published findings might be spurious" (p. 984). In this article the Sommers' describe how a dramatic, but inadequately documented and unreplicated study on the effects of early intervention on intelligence managed to seep into textbooks in two major areas of psychology.

The following complaint will, I think, be familiar to everyone who has followed research in parapsychology: "The most obvious deficit in the literature reviewed is the rarity with which experiments from one laboratory are replicated in another. However, there is also some persisting doubt as to the consistency of the effects found within a laboratory. Anyone who has worked in this field is aware of this problem . . . It is, therefore, imperative that researchers report sufficient replications with adequate statistics to substantiate their responses."

While similar evaluations can be found in the parapsychological writings of several of the participants at this conference, this particular observation represents one expert's evaluation of the status of research on the neurochemistry of learning and memory (Dunn, 1980).

My intention is not to minimize the repeatability problem in our own field, there is one, but rather to put it in perspective. Too often discussions of replicability in parapsychology have treated psi research as though it existed in a vacuum. But, of course, it doesn't; other areas have repeatability problems too.

### *Belief, Expectation and Experimenter Effects*

Is there something unique about the replicability problem in parapsychology that distinguishes it from replicability problems in other areas? Two factors that have sometimes been cited in this connection are belief and experimenter effect. The existence of an experimenter effect is undeniable. Some experimenters find psi effects in their experiments while others do not. Some "successful" experimenters appear to be more "successful" than others.

There is, as many of you know, a large literature on experimenter expectancy effects in psychological research which appears to show that nonverbal communication to subjects of the experimenter's expectations can bias subjects' responses in a variety of experimental tasks (Rosenthal and Rubin, 1978). A number of studies, for example, suggest that the experimenter's tone of voice is sufficient to convey her expectations to subjects. Consider a two-stage study by Adair and Epstein (1968). In stage I, experimenters were led to expect either high or low ratings from their subjects and the experimenters obtained results significantly in the direction of their expectations. In stage II, there were no experimenters at all. Instead, tape recorded voices of the experimenters instructing subjects in stage I were played for new groups of subjects. The results of stage II showed that the effects of experimenters' expectations were communicated as effectively by tape as they had been in the live subject/experimenter interactions of stage I.

Rosenthal and Rubin summarized 345 studies of experimenter expectancy effects in eight areas of research, ranging from inkblot tests and psychological interviews to learning and reaction time experiments. They found that approximately 35 percent of these studies were significant at the 5 percent level or lower.

The sheep-goat effect (Palmer, 1971), among other parapsychological research findings, indicates that the psi performance of subjects is modulated to some extent by their belief/disbelief in ESP, and it is not unreasonable to suppose that "sheep" and "goat" experimenters may communicate somewhat different expectations of success to their subjects.

Another area in which belief and experimenter effects exert a powerful influence on experimental outcomes is placebo research. Placebo research demonstrates the powerful effects of belief and disbelief on the treatment of a wide variety of physical and psychological disorders. Since the placebo is pharmaceutically inert, the treatment is symbolic. Expectation of success, on the part of both physician and patient, appears to be the most significant factor in successful placebo treatment. Like our notoriously "elusive" psi effects, placebo efficacy is highly variable, and this variability exists across physicians and hospitals, just as in psi research variability exists across experimenters and laboratories. The following brief sketch is drawn from a recent review of placebo research by a medical anthropologist at the University of Michigan, Daniel Moerman.

Since the seminal paper on placebo effects by Beecher (1955), many studies have appeared indicating that placebo relief of pain is,

on the average, effective in about 35 percent of cases. Emphasis on mean effectiveness, Moerman points out, has masked the huge variability of effectiveness. Placebo healing rates in the treatment of ulcers, for example, have varied in recent studies from 8–83 percent (Bodemar and Walan, 1976; Blackwood, Maudgal and Pickard, 1976; Schuerer, Witzel, Halter et al., 1977).

Attempts to account for the dramatic variability of placebo effects have focused on four variables: personality characteristics of the patient, the nature of the illness, the set and setting in which the placebo is administered and the physician.

Most research attempting to directly clarify the role of the placebo has involved patient characteristics. Typically patients are divided into two groups—responders and non-responders, which were compared on a number of personality measures. This approach has by and large been fruitless, according to Moerman, who concludes that "the least significant variable in the equation is the personality of the patient" (p. 259).

Placebos seem to provide effective treatment for a diverse range of illnesses, including pain, rheumatoid arthritis, warts, acne, wound healing, angina, hypertension, anxiety and depression. "That pain, immune mechanisms, and coronary vasospasm are all EQUALLY accessible to symbolic manipulation seems unlikely," Moerman says, adding that, ". . . published data exhibit such wide variations in placebo effect rates WITHIN syndromes as to prevent useful comparisons BETWEEN them . . ."

While patient personality factors and the nature of their disorder have not illuminated the variability of placebo response, one very powerful factor can be identified: the patient's expectation of success. Several studies have shown that two placebos are more effective than one (Rickels, et al., 1970) and that patients receiving different placebos in three consecutive two-week periods improved more than a group receiving one placebo for a six-week period (Rickels, et al., 1963). Other studies indicate variable placebo effectiveness as a function of the color of medication. In one British study, for example, medical students were told they were testing either stimulants or sedatives and pink placebos were found to act as stimulants while blue placebos acted as sedatives (Blackwell, Bloonfield and Buncher, 1972).

Placebo effectiveness has been found to correlate with the reputation of effectiveness of the drug for which it is substituted. In a double-blind psychiatric study, a 24 percent placebo response rate was obtained when patients believed they were taking a mild tran-

quilizer, a 35 percent rate when compared to moderate doses of a stronger tranquilizer and 76 percent placebo response when patients believed they were taking still-stronger tranquilizers (Lowinger and Dobie, 1969).

For present purposes, the most interesting parallel between placebo research and psi research is the existence of a strong physician effect in placebo studies. Two double-blind studies of placebos in ulcer treatment showed substantial placebo effectiveness compared with an untreated control group (Sarles, Camatte and Sahel, 1977). However, large and significant differences in placebo outcomes were found in the groups treated by different physicians. Number of days of ulcer pain was the measure of treatment effectiveness. For an untreated control group, the mean was 19.5 days. For the groups treated with placebos, the results varied as a function of physician: 12 days of pain for patients treated by one physician, seven days for the patients of two other physicians and 3.5 days for a fourth physician. In other words, one physician was three times more effective than another in alleviating pain using inert treatments.

Reviewing the history of several now discredited treatments for angina, Benson and McCallie (1979) note the dramatic difference in effectiveness rates depending on whether the physicians were enthusiastic or skeptical regarding the value of the medication: "the initial 70 to 90 percent effectiveness in the enthusiasts' reports decreases to 30 to 40 percent 'baseline' placebo effectiveness in the skeptics' reports." The notion among physicians that drug effectiveness is historically transitory is reflected in the well-known adage, "Treat as many patients as possible with new drugs while they still have the power to heal."

Similar differences have been reported in studies conducted in different hospitals. Moerman reports that the same antacid treatment for ulcer had a 79 percent effectiveness rate in one hospital and a 17 percent rate in another. Finally, considering the differential success rates of American, British and European psi studies, it is interesting to note that several authors have estimated placebo healing rates for ulcer to be as much as twice as high in the United States as they are in Europe and Great Britain (Hirschowitz, 1977; Gudjonsson and Shapiro, 1978).

*Measuring Replication Rates*

Let us turn now to a consideration of how we can improve our own situation. I believe parapsychology has already taken a very

important step in the right direction. I refer to the fact that unlike the psychological journals, replications in parapsychology are encouraged and highly valued. Indeed, I believe it is not an exaggeration to say that experimental findings in psi research are never taken very seriously among research workers in the field until there has been at least some in-house replication. Recognizing the importance of negative results in assessing the status of psi research, the Parapsychological Association Council in 1975 adopted a policy opposing the selective reporting of positive results. Nonsignificant findings are routinely reported at PA meetings and in PA-affiliated journals.

We have tended to pay too much attention to arbitrary significance levels and not enough to estimates of effect size and variability of effect. Astronomical p-values do not necessarily imply strong effects. A fast random number generator PK experiment with a million trials and a cumulative deviation from chance of two-tenths of one percent would be associated with odds of nearly a million to one. Yet our habits of thought have generally led us to be more impressed by small p-values than large effects. The following statement by Professor Hansel (1980) is not atypical: "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, p. 298).

Now consider a psi Ganzfeld experiment involving 30 trials with a probability of a hit on each trial of ¼ and 14 hits. That would be a success rate of 47 percent, nearly twice the expected chance rate of 25 percent, and it would be statistically significant at the .01 level. According to Hansel, if the result of this experiment involved some extrachance factor, such as ESP, we should be able to confirm this fact simply by collecting two additional sets of data, each of which should (Hansel assumes) yield the same result, so the overall probability for all three data sets would be one in a million. An important flaw in Professor Hansel's reasoning is that it assumes a constant effect size from one experimental sample to another. It ignores the fact that the experiment's success rate of 47 percent is really only one estimate of the effect size. The 95 percent confidence interval for our hypothetical Ganzfeld study is plus or minus 15 percent; if there is a real effect, its actual population mean could be anywhere between 32 percent and 62 percent. The practical consequence is that

replications of the experiment (each with n = 30 trials and .01 significance level) will frequently fail to detect a real effect.

This point was made earlier in a review of the sheep-goat effect by John Palmer (1971). Assuming a small, but real sheep-goat effect whose magnitude is on the order of Schmeidler's group experiments, Palmer (1971) calculated a theoretical sampling distribution of mean differences between sheep and goats. This led to three predictions: (a) sheep should score higher than goats in approximately 84 percent of the experiments, (b) statistically significant sheep-goat effects should occur in only about 16 percent of the experiments and (c) less than 1 percent of the experiments should show significant reversals (i.e., goats scoring significantly higher than sheep). Reviewing the available sheep-goat studies, Palmer found that his predictions were in reasonable agreement with the experimental findings: (a) sheep scored higher than goats in 76 percent of experiments, (b) significant sheep-goat effects occurred in 35 percent of the experiments and (c) there were no significant reversals.

An important problem in the assessment of replicability rates in groups of studies is that quantitative results are often not presented in adequate detail. Reports of nonsignificant studies in particular often fail to provide sufficient quantitative data. Frequently authors seem to feel that their failure to reach the magic .05 level says all that needs to be said about their results and too often the only numerical information given is in the form "p = nonsignificant." Regardless of the outcome of the study, the results should be reported in sufficient detail that an interested reader can calculate the effect size and determine whether or not the result lies within the confidence limits of the hypothesized effect.

A movement has emerged within psychology in the last few years that is directed toward the quantitative integration of entire research domains. "Meta-analysis," as it is called, employs statistical analysis across groups of studies using the studies' outcomes as dependent variables and their designs, procedures and sampling parameters as the independent variables. The goal of meta-analysis is to quantify and estimate the strength of relationships between study outcomes and factors in the experimental procedures. As described by one of the principal pioneers in this movement, it is ". . . nothing more than the attitude of data analysis applied to quantitative summaries of individual experiments. By recording the properties of studies and their findings in quantitative terms, the meta-analysis of research invites one who would integrate numerous and diverse findings to apply the full power of statistical methods to the task. Thus it is not

a technique; rather it is a perspective that uses many techniques of measurement and statistical analysis" (Glass, McGaw and Smith, 1981, p. 21).

Unlike traditional narrative style literature surveys, meta-analysis does not prejudge research findings in terms of research quality, but, rather, seeks empirical evaluation of suspected weaknesses or flaws through statistical comparison of groups of studies with and without the suspected flaws. Recent examples in parapsychology include the evaluation of potential flaws in psi Ganzfeld research (Hyman, 1983; Honorton, 1983) and in studies of the effect of hypnotic induction procedures on psi performance (Schechter, in press). Since Professor Hyman is in the process of revising his evaluation of the Ganzfeld research, I will restrict myself to a brief description of Schechter's review of the hypnosis work.

Schechter located 25 published comparisons of ESP performance with and without hypnotic induction in 20 papers by a dozen researchers in ten different laboratories. Five of the studies were omitted from the analysis because induction/control comparisons could not be unambiguously interpreted. Of the 20 studies used in the analysis, 16 studies—80 percent of the total—showed higher ESP scores following hypnotic induction than in the control condition, indicating a significantly consistent directional effect ($p < .006$). Whereas by chance only one of the 20 studies would be expected to show a significant difference at the 5 percent significance level, seven of the studies or 35 percent showed significant differences favoring the hypnotic induction condition, a result which cannot reasonably be attributed to chance ($p < .000034$). None of the studies showed significantly higher scoring in the control condition. The overall pattern of results therefore supports the hypothesis that higher ESP scores occur in the hypnotic induction condition.

Can these results be reasonably attributed to selective reporting of positive findings? Six of the seven independently significant studies were reported prior to 1975, when the PA announced its policy opposing selective reporting. While there is no way to be certain, there are two considerations that militate strongly against this possibility. The first consideration is that nonsignificant findings were in fact reported in more than half of these studies. The second consideration involves estimation of how many additional negative studies would be required to raise the cumulative probability of the known studies to $p = .05$. Using the method described by Rosenthal (1979) for assessing tolerance for null results, Schechter estimated that it would take 97 unreported studies with hypnosis/control differences

averaging zero to raise the cumulative probability of the differences found in the 20 reported studies to $p = .05$. The selective reporting hypothesis is viable only if we are willing to assume that for every reported hypnosis/control comparison there are six unreported comparisons showing no difference between conditions.

Schechter's next step involved comparison of the studies' outcomes in relation to identifiable weaknesses in their designs and procedures. Six design or procedural problems could be assessed from the reports: controls against sensory cues, method of randomization, controls for recording and checking errors, the number of participants and number of trials/condition. No significant relationships were found between study outcomes and the presence or absence of design problems in any of these six areas individually or collectively. The correlation between the cumulative number of flaws and success of study was very close to zero.

Schechter concluded that ESP performance appears to be reliably stronger following hypnotic induction. Unfortunately, the current batch of studies tells us only that induction facilitates ESP performance; it does not tell us how. What are the controlling variables? Is there something special about the hypnotic state? Does the ritual of induction function as a placebo, increasing experimenter/subject expectations of success? These questions and a myriad of others set the agenda for a new and more finely-focused generation of studies.

Meta-analysis offers powerful new tools that can help us integrate whole areas of research, test hypotheses about process and empirically evaluate issues concerning research quality. It provides a structure for drawing generalized and empirically-anchored conclusions from our data. But in order to maximize the potential of meta-analysis, we must develop more uniform standards for reporting the individual research studies which supply the raw data for meta-analysis.

A number of research areas in parapsychology have been active for a decade or more. Ganzfeld research and random generator PK research are two examples. By my last count, 48 individual psi Ganzfeld studies have been reported in the last ten years. I haven't systematically surveyed the RNG area recently, but a reasonable estimate is that there must be at least 65 studies in that area by now. Surely we have learned enough from our successes and failures in these areas to agree on some minimal standards for reporting.

Yet research reports in both areas remain uneven and lack uniformity of description. As I mentioned earlier, many unsuccessful replication efforts provide no quantitative information at all beyond the studies' failure to reach statistical significance. Process-oriented

studies, while often devoting considerable space to descriptions of the experimenter's hypotheses and manipulations, sometimes fail to describe such elementary information as what instructions subjects were given regarding the psi task! Questionnaires and other instruments used as predictors are frequently reported without descriptive statistics to assess sampling characteristics that might illuminate interlaboratory differences.

To some extent, of course, what we decide is important enough to include in the experimental report is a matter of trial-and-error and we cannot report everything. Take a seemingly trivial example: how detailed a description need be given in the METHODS section of a report regarding the application of ping-pong balls over the subject's eyes in a Ganzfeld study? We once had a visitor at Maimonides who reacted very negatively when we invited her to participate in a Ganzfeld study. She had recently done a Ganzfeld session in another laboratory and clearly regarded the Ganzfeld as a form of torture. It seems the experimenter had devised a novel method to insure that the subject complied with instructions to keep her eyes open during the session: he taped them open. After half an hour in Ganzfeld, the poor subject's eyes were swimming in a sea of tears. Harry Collins was right, the transfer of knowledge is indeed a capricious process!

Nonetheless, there are quite a number of things that I think we could agree should be routinely included in experimental reports, which should increase the likelihood of successful replication. It might be appropriate for the PA to once again take the lead and to commission a task force in areas that have demonstrated promising success rates over some reasonable period of time. The task force would be composed of experienced workers in the area consisting of both "successful" and "unsuccessful" investigators and their job would be to develop reporting standards that could serve as guidelines for the editors of PA-affiliated journals, providing some degree of uniformity of reporting in a given area. The guidelines would provide standards for reporting procedural conditions, such as randomization and control of sensory cues, as well as potential moderator variables.

Whether through the PA or another mechanism, I believe that assessment of progress in parapsychology will require us to develop criteria for evaluating aggregates of studies just as in the past we have developed criteria for evaluating the significance of individual studies. The several diverse perspectives offered at this conference provide an excellent basis for initiating consensus-forming dialogue. Shall we begin?

## BIBLIOGRAPHY


Adair, J. G. and Epstein, J. S. "Verbal cues in the mediation of experimenter bias." *Psychological Reports*, 1968, *22*, 1045–1053.

Barber, T. X. "Expecting expectancy effects: biased data analyses and failure to exclude alternative interpretations in experimenter expectancy research." *The Behavioral and Brain Sciences*, 1978, *3*, 388–390.

Beecher, H. K. "The powerful placebo." *Journal of the American Medical Association*, 1955, *159*, 1602–1606.

Benson, H. and McCallie, D. P. "Angina pectoris and the placebo effect." *New England Journal of Medicine*, 1979, *300*, 1424–1429.

Blackwell, W. S., Bloonfield, S. S. and Buncher, C. R. "Demonstration to medical students of placebo responses and non-drug factors." *Lancet*, 1972, *1*, 1279.

Blackwood, W. S., Maudgal, D. P., Pickard, R. G., et al. "Cimetidine in duodenal ulcer: Controlled trial." *Lancet*, 1976, *2*, 174–176.

Bodemar, G. and Walan. "Cimetidine in the treatment of active duodenal and prepyloric ulcers." *Lancet*, 1976, *2*, 161–164.

Bozarth, J. D. and Roberts, R. R. "Signifying significant significance." *American Psychologist*, 1972, *27*, 774–775.

Collins, H. "The TEA Set: Tacit Knowledge and Scientific Networks." *Science Studies*, 1974, *4*, 165–186.

Collins, H. "Science and the Rule of Replicability: A Sociological Study of Scientific Method," paper presented at a symposium, "Replicability and Experimenter Influence," at the Annual Meeting of the American Association for the Advancement of Science, Washington, D.C., February, 1978.

Dunn, A. J. "Neurochemistry of learning and memory: an evaluation of recent data." *Annual Review of Psychology*, 1980, *31*, 343–390.

Glass, G. V., McGaw, B. and Smith, M. L. *Meta-analysis in Social Research*. Beverly Hills, CA: Sage, 1981.

Gudjonsson, B. and Shapiro, H. M. "Response to placebos in ulcer disease." *American Journal of Medicine*, 1978, *65*, 399–402.

Hansel, C. E. M. *ESP and Parapsychology: A Critical Re-evaluation*. Buffalo, NY: Prometheus, 1980.

Hirschowitz, B. "Histamine H-2 receptor antagonists." *Annals of International Medicine*, 1977, *87*, 373–375.

Honorton, C. "Response to Hyman's critique of psi ganzfeld studies." In W. G. Roll, J. Beloff and R. White (Eds.), *Research in Parapsychology 1982*. Metuchen, NJ: Scarecrow Press, 1983.

Hyman, R. "Does the ganzfeld experiment answer the critics' objections?" In W. G. Roll, J. Beloff and R. White (Eds.), *Research in Parapsychology 1982*. Metuchen, NJ: Scarecrow Press, 1983.

Levine, J. D., Gordon, N. C. and Fields, H. L. "The mechanisms of placebo analgesia." *Lancet*, 1978, *2*, 654–657.

Littma, A., Welch, R. and Fruin, C., et al. "Controlled trials of aluminum hydroxide gels for peptic ulcer." *Gastroenterology*, 1977, *73*, 6.

Lowinger, P. and Dobie, S. "What makes the placebo work? A study of placebo response rates." *Archives of General Psychiatry*, 1969, *20*, 84–88.

Moerman, D. E. "Edible symbols: The effectiveness of placebos." In Thomas A. Sebeok and Robert Rosenthal (Eds.), *The Clever Hans Phenomenon. Annals of the New York Academy of Sciences*, Volume 364, New York: NYAS, 1981, 256–268.

Palmer, J. "Scoring in ESP tests as a function of belief in ESP. Part I.: The sheep-goat effect." *Journal of the American Society for Psychical Research*, 1971, *65*, 373–408.

Rickels, K., Baum, C. and Fales, K. "Evaluation of placebo responses in psychiatric outpatients under two experimental conditions." *Neuropsychopharmacology*, 1963, *3*, 80–84.

Rickels, K., Hesbacker, P. T., Weise, C. C., et al. "Pills and improvement: A study of placebo response in psychoneurotic outpatients." *Psychopharmacologia*, 1970, *16*, 318–328.

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

Rosenthal, R. and Rubin, D. B. "Interpersonal expectancy effects: the first 345 studies." *The Behavioral and Brain Sciences*, 1978, *3*, 377–415.

Sarles, H., Camatte, R. and Sahel, J. "A study of the variations on the response regarding duodenal ulcer when treated with placebo by different investigators." *Digestion*, 1977, *16*, 289–292.

Schechter, E. "Hypnotic induction vs control conditions: illustrating an approach to the evaluation of replicability in parapsychological data." In R. White and R. Broughton (Eds.), *Research in Parapsychology 1983*. Metuchen, NJ: Scarecrow Press, 1984.

Sommer, R. and Sommer, B. A. "Mystery in Milwaukee." *American Psychologist*, 1983, *38*, 982–985.

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

## *DISCUSSION*

RAO: I think you have some good ideas. I think your talk is a fitting conclusion to this program. What I am going to say will more or less reinforce what you have said. I have two or three specific points.

With regard to the meta-analysis that you have recommended, it might be worthwhile to remember at this point that the first psi meta-analysis in any behavioral field or perhaps in any field was done in parapsychology. This was the analysis in *Extra-Sensory Perception After 60 Years* by Rhine, Pratt, etc., and it was acknowledged in the most recent issue of the *Psychological Bulletin* where the lead article has to do with meta-analysis. So we are going back and I am glad that we are going back in the right direction.

As to the placebo effect, I think that both of us seem to be on the same wavelength. Recently I did a review of the placebo literature for a lecture on medical implications of parapsychology. I happened to read some of the original papers. I was impressed by two things as I read these papers. One, their methodological sloppiness, a perpetual commitment to small numbers and the total carelessness of the controls. And second, as you wade through this jungle of literature you cannot fail to be impressed by the fact that, while there seems to be something in the placebo effect, there also seems to be room for psi. This was particularly true in one study, along the lines that you have mentioned, where patients of two physicians

were administered placebos and active drugs. In one the physician himself was not directly in contact with the subject. I understand, if I can trust these results, that in the case of the physician who did not believe in the placebo, even though he had nothing to do with the administration of it, the placebo did not work. On the other hand, in the case of the patients of the physician who did believe in the efficacy of the placebo, it worked, even though the experiment was totally blind. Here again, I think that psi might be going on in some of these processes, so I feel that we are on the right track. There are many borderline areas, such as experimenter expectancy where there might be commonalities between psi research and placebo research.

HONORTON: I spoke recently with one of the early pioneers in placebo research. He said that he thought for many years that there was probably a parapsychological component to placebo effectiveness. We are going to get together in the next few weeks. I was in a major medical center for 12 years and it never occurred to me to do any research on this at the time. Now we are in a research park out in Princeton, New Jersey and we don't really have access to medical facilities. But this could be a very exciting area of research because it would have all kinds of implications for theory. The main point, however, is that here is something where there is clearly an effect of psychological state on human physical functioning. It shows many of the same characteristics, at least according to this review of the literature, as do our own—experimenter effects, the equivalent of laboratory differences, declines over time and so on.

SCHECHTER: I would like to make two comments. One is a relatively minor emendation to Chuck's review of my beginning meta-analysis of the research on hypnotic induction versus control conditions. Chuck did make sure I was satisfied with the way he'd written the summary. But in listening to it now, I caught a turn of phrase that I missed before. Chuck said that there were no significant relationships in the various breakdowns of procedural problems versus outcomes. I want to remind those who don't remember the original paper that I only did significance testing on the measure that combined the various flaws: The differences in the breakdowns of individual flaws were minor, but the numbers were too small to justify significance testing.

The other comment is to back up something that a number of people have said in the past couple of days that I feel rather strongly about—and that is to put in a plea to people writing up reports to please say what you did! In a number of the reports I reviewed, all

that was said about the induction procedure was that the subjects were hypnotized. Some writers went into great detail, or provided references to procedures. Many said next to nothing. There was no way of comparing the adequacy of the procedures. Was the depth and adequacy of the induction monitored? Most reports did not mention monitoring, and some of the reports which did mention it did not indicate how the monitoring was done. It became a very frustrating job to try to pull this material together into a meta-analysis. I see Chuck nodding his head; I know we have gone through the same thing with the Ganzfeld literature, and I am sure that the rest of you at one time or another, have stumbled across the same problem. It gets to the point where it hurts—we need to pay some attention to this kind of thing.

HONORTON: It is not as though the parapsychological journals have so much good material coming in for each quarterly issue that it is necessary to reduce the length of experimental reports, particularly in areas such as Ganzfeld and RNG and bio-PK and others that are undergoing continual evaluation and assessment. It is absolutely essential that we start reporting things in detail. It is easy to say that, but from the example that I gave of how to put ping-pong balls on the subject's eyes it is quite clear that we do not always know what it is necessary to report until we have detected some kind of problem like this. I think that the Parapsychological Association should take the initiative and set up a task force of this type. It would be a relatively easy thing to do and should be composed of successful and unsuccessful experimenters, so that we cover the entire terrain of concerns of methodology and potential moderator variables. It should be very helpful to the editors of the journals in providing guidelines for evaluating manuscripts.

RAO: As an editor of a journal in the field, I feel that space restrictions have never prevailed in cutting down procedural details that any author had given. We scrupulously follow the comments of the referees as to any further detail they might want to be given in the journal. So if the journal articles in recent years have failed to provide the necessary information regarding procedure or an evaluation it is not so much the fault of the author alone, but the fault of the people who read and refereed them. The people who refereed those articles are the leaders in the field, most often people who are making these criticisms, too.

HONORTON: I would like to respond to that briefly, because I think that is true. *Research in Parapsychology*, of course, is a different situation. I myself have been frustrated by seeing a Douglas Stokes

review of an *RIP* that criticized a study of mine, an RNG study, for example, for relying on oscillating the target as a control rather than using control trials, when in fact we had control trials and reported control data, but that was edited out of the final version.

BERGER: I would like to address the issue of standardization, specifically as it applies to the Ganzfeld and the random number generator work. Although there is a body of data being cumulated, one still has great difficulty attempting a meta-analysis of these areas, because you cannot be sure from the published reports whether studies are comparable on certain critical factors. So I propose that standardizing of reporting criteria is a necessary first step. I think it needs to be stressed that meta-analysis is not possible unless complete results, including non-significant analyses, are reported. This means that if ten experimenters each do an identical experiment, each fails to reach significance and each reports "non-significance" instead of the exact results, a significant finding (such as that nine out of ten were in the predicted direction) would be completely overlooked.

We have also discussed, among ourselves and with other laboratories, the idea of beginning to standardize procedures and methods. In the RNG work, for example, that might mean using the same type of RNG, agreeing on sampling frequencies, or at least starting to do experiments that enable us to later compare results across laboratories. The minimum that should be done is the reporting of these details so at some later point in time they can be systematically analyzed.

HONORTON: I have frequently felt in looking at the RNG literature that, except for Helmut Schmidt, hardly anyone describes what the random number generator is. It might be that Helmut has a TEA laser and John Beloff maybe has a piece of hardware that is not a TEA laser, but looks like one except it doesn't function in the same way. This may be absurd but we really don't know. The hardware should be at least described uniformly. To the degree that it is reasonable and practical to do so, some degree of standardization of hardware is obviously a good idea because there at least we can eliminate that as a likely difference between laboratories.

BLACKMORE: I have been struck by the fact that you, in common with other people today, have kept on picking out the Ganzfeld work and the REG work as being the most worthy of note, the most hopeful for the future. I am just looking at William Braud's table and it is interesting to note that if you look at the percentage of replication of studies these two are the lowest—with REG at 35 percent and Ganzfeld at 48 percent, whereas the early ESP card

guessing which no one has mentioned at all, comes out at 82 percent. I don't know what this tells us, but it may be that replication is simply not what we are doing when we have come to judgments about which things are promising for the future. I wondered if you had any other thoughts about whether it tells us anything?

HONORTON: I noticed that also. I noticed it some time earlier. I don't know with the early ESP card guessing—that bothers me a little because that doesn't seem right to me. I was going by the references in *Extra-Sensory Perception After 60 Years.* My assumption was that during that period of methodological controversy it was not likely that there would be a lot of nonsignificant studies withheld from publication, because it was a big issue at that time and nonsignificant studies were getting into the psychological journals. But I don't know. Obviously, Ganzfeld has recently gone through a much more refined process of evaluation and examination and elimination of studies that were reported to be significant that were not when you corrected for multiple analysis. I don't think that most of these other areas have yet gone through that pruning process. So I don't know really how comparable all these areas are.

BLACKMORE: And yet, we *are* going on something when we pick on those two.

HONORTON: Well, I pick on them simply because I am familiar with them and I have been working on them.

STANFORD: I agree with what Chuck just said. There is another consideration that bears upon the comparability question. One is the differing number of trials of studies that are done and what is constituted by one of these numbers—the unit 1 that goes into a number like 28 for remote viewing studies for example. I really think it is almost meaningless to do so. Many of them have not been filtered for the kinds of criteria that Chuck alluded to and the facts that he had gone over in the Ganzfeld review.

HONORTON: I want to emphasize this point. In fact, earlier this morning, Dr. Rao mentioned the Rosenthal-Rubin paper and the fact that they have only a 35 percent replication rate for the experimenter expectancy effect and that is lower than in many areas in parapsychology. That is true; that is 35 percent of 345 studies. And with the RNG work here we have some 214 data sets. Having gone over the May, Humphrey and Hubbard paper, I don't know how they classified data sets here because there are sometimes different conditions within an experiment and so on. But 35 percent out of 214 is much more impressive than 48 percent of 48, obviously.

SCHMIDT: A short comment on the random generator. I haven't

yet studied the summary carefully, but my impression was that when random generators became fashionable many outsiders with very little experience in parapsychology jumped at this psychologically attractive approach. That might explain much of the relatively low success rate.

HONORTON: I would have to dispute that. Some of the early nonsignificant Ganzfeld studies were pretty poorly done.

SCHECHTER: A short note on this topic of the numbers. Has anybody else noticed the ubiquitous 35 percent? The placebo work which has also a large data base that Chuck discussed also comes in at around 35 percent. Curious, is the most I will say at this point.

BELOFF: I would just like to make a brief reference again to the placebo research. Before I heard Chuck Honorton's paper I really had no idea that so much work had been done on the placebo effect, but now that I am enlightened about this I am very intrigued because, after all, if one asks oneself what is happening in a placebo effect of this kind, it brings one back again to something very like a psi effect. Ordinarily, we don't want to include it as anything paranormal because we say the patient expects to get better, therefore he does get better. In that way of putting it we don't think that anything paranormal has happened, but when you think about it there is very little explanation in any normal physical sense of why an expectation should generate the kind of physiological processes that are necessary for healing. It looks much more as if some kind of PK effect is being exercised on one's organism in this case. But because it is intraorganismic one can't prove that there is no normal explanation. It is hypothetically always there. But as parapsychologists we take the psi hypothesis seriously. We ought not to feel so surprised that the placebo research should reveal results of this kind and we should seek to relate it to our knowledge of the way psi works in the kinds of studies we carry out.

HONORTON: Yes, I agree completely with that. One problem in thinking about how you go about doing a parapsychological study involving the placebo effect is that it is quite clear from Moerman's review that you couldn't separate the patient and the physician in order to do the study. It is the expectation of both that really seems to be important. Now, one possibility would be if we could get the cooperation of a medical school or medical centers doing drug studies in this area. We would set up a microcomputer or one of Helmut's devices somewhere and perhaps we could find a way to allow medical doctors to test their patient's PK without being run out of their

profession. That might be one way to do that. I would expect that a better psi task for this, in fact, one that would be almost ideal, would be William Braud's allobiofeedback procedure because that itself involves a very similar kind of process. The ultimate irony to me in going through the placebo literature—it is fascinating stuff, extremely important, whether psi is involved or not—is that almost all of the research has not been oriented toward the placebo, it has been oriented toward removing the placebo. What they want to find out in these studies is how does the drug that is active work and that is the reason for the small samples in most of the studies. The probable reason for the sloppy methodology in many of them is that the placebo effect is not the primary topic of interest. But I agree. I think it is very important that we try and develop some research that will allow us to do some correlations between placebo effectiveness and psi. Because, if there is a correlation there, then it would have extremely important theoretical implications.