REPLICATION IN CONVENTIONAL AND CONTROVERSIAL SCIENCES

K. RAMAKRISHNA RAO

In normal science, repetition of an experiment or the possibility of it is not a matter of primary importance. If it were, astronomy, for example, would not be regarded as a science because planetary positions do not repeat themselves. I browsed through a dozen highly regarded books on philosophy of science and only in two of them did I find an entry in the index for repetition, repeatability, or replication. Even in the two that I did find, the discussion of the problem of replication was rather peripheral and limited to a few lines.

This state of affairs is understandable because, when experiments are repeated, it is for reasons such as (a) improving experimental techniques, (b) accumulating more data and thereby increasing their accuracy and generalizability and (c) checking on the competence of the experimenter. But in controversial sciences and when anomalous claims are made, replication takes on greater importance because questions that are of secondary concern in normal science are now raised to a level of primary importance. In the minds of skeptics, replication then becomes the *sine qua non* of good science, a criterion that is used to distinguish the genuine from the spurious. Existential questions are raised and answered on the grounds that only replicable effects are genuine and that only repeatable phenomena are real.

Accordingly, psi phenomena are often rejected by skeptics for the alleged reason that they are not replicable. "If there is one common and basic feature of experimental science," write Moss and Butler (1978), "it is the possibility of the reproduction of findings by independent investigators. . . . Replication by a qualified nonsympathetic observer is the only guard against results which may have been contaminated by conscious or unconscious bias" (p. 1067, 1068). They argue that a number of psychologists failed to obtain significant results in psi experiments. Therefore, parapsychological phenomena must be presumed to be nonexistent.

The preceding argument makes several assumptions. First, replicability is the basic feature of all experimental science. Second, to rule out conscious or unconscious bias, replication must be made by a "qualified nonsympathetic" observer. A corollary of this is that scientific observations are completely unbiased and impersonal. Third, failure to replicate is sufficient reason to reject a phenomenon as spurious.

It is easy to see that all these assumptions are questionable. Replication is not a common practice in science. In hard science, one rarely comes across publications that are mere repetitions of previous experiments. Even in behavioral sciences, replication studies are relegated to a secondary and low status. Only when a claim is controversial is the question of replication raised.

The demand for replication by a qualified nonsympathetic observer is unreasonable in the extreme. First of all, in some areas where observer sensitivity is an important qualification for eliciting a phenomenon, a "nonsympathetic observer" may not satisfy an essential qualification. Thus, "qualified nonsympathetic observer" may be a contradiction in terms. Again, the notion of unbiased observer is problematic. Much of contemporary philosophy of science concedes a crucial role for the observer. As N. R. Hanson (1958) asks us, imagine that Johannes Kepler and Tycho Brahe are both watching the dawn in the east. What do they see? While Kepler sees the rim of the earth drop away, Tycho sees the sun rise up. Yet, they have essentially the same visual stimulation. Michael Polanyi (1950, 1967) has emphasized the tacit dimension in *knowing* and the personal aspects involved in the acquisition of knowledge. The perceptual process, as Polanyi points out, consists in the tacit integration of perceptual clues into comprehensive entities. Perception is, in a sense, the transposition of feelings. The way we see an object is mainly determined by our awareness of certain events in our bodies, which are not themselves observable.

Even if we accept the traditional view of science and its foundationist epistemology (Carnap, 1956) and believe in the possibility of a preinterpreted "given," there are procedures for shielding the experimental results from being contaminated by observer bias. Double-blind procedures, for example, are meant precisely for this purpose. As I pointed out elsewhere (Rao, 1979): "The argument that a skeptical experimenter is the only person whose replication of an ESP experiment is valid is untenable on several grounds. (a) A skeptical experimenter may be just as strongly biased against ESP as the believer may be in the opposite direction. By the same argument,

the skeptic's findings when null or negative would carry no greater credibility than the positive findings of the believer. (b) The nature of psi may indeed be such that a negatively motivated experimenter would interfere with its occurrence. This possibility is not one that is unique to parapsychological phenomena. For example, certain experimenters do not, for a good reason, obtain Rosenthal's experimenter expectancy effects (Rosenthal, 1976). (c) When a skeptic obtains significant psi results, he ceases to be a skeptic. Since he is now a 'believer,' his positive results, by this logic, would not be expected to carry any weight with other skeptics. (d) There is no guarantee that a person who is skeptical of psi would employ the correct experimental procedures, would draw only legitimate inferences, or would be honest" (p. 419).

As mentioned earlier, replication has little role to play in normal science. Even in controversial areas where it is demanded by skeptics, replication is not a demarcating criterion to distinguish the genuine from the spurious.

As Collins has argued, replication is a social process, and there is sometimes room for disagreement whether or not a finding is replicated. If a student in a chemistry laboratory does not obtain the results he is supposed to, we conclude that the student did not conduct the experiment properly, because we have already accepted the original result as an established fact. In a case where we have not accepted the results of a study as genuine, we are more likely to interpret a failure to replicate as a refutation of the original claim, rather than question our own ability to carry out an exact replication of the original study. Thus, our interpretation of the results of an attempted replication depends strongly on our prior notions about the credibility of the original finding. Consequently, to regard replication as the criterion for demarcating the genuine and the spurious is to beg the question.

The crucial point is that a basic condition for replication is that it be an *exact copy* of the original. But the notion of achieving an *exact copy* is itself problematic, inasmuch as some of the knowledge involved may be "tacit" and not subject to clear articulation, more like acquiring a skill than communicating a formula. Collins (1978) illustrates this point by referring to seven attempts to build TEA lasers after the details of inventing this device were made public in 1970. As he put it, "where scientists tried to build a laser based 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 the sources of help" (p. 9).

So then, the possibility that tacit knowledge may be required to replicate a finding successfully leaves open the question of whether a replication attempt is an exact copy of the original. The proponents of the original finding could argue, when the replication fails, that it is not an *exact copy*, whereas the opponents may insist that it is. Therefore, Collins (1978) observes: "in controversial areas replicability does not work as a demarcation criterion for the genuine and the spurious, but rather, repeatability is a notion that is attributed to what are considered genuine phenomena" (p. 19).

Replication and New Realist Philosophy of Science

Let me expand this point a bit further by referring to the new realist philosophy of science as expounded by Rom Harré (1970) and Roy Bhaskar (1975, 1982), among others. The new realism is the third force in philosophy of science. The traditionalists emphasize hypothesis testing, by verification or falsification, as central to science. They believe that hypotheses are to be tested against "facts" and that the truth of a theory is the correspondence between its constructs and observations. Apart from the fact that the distinction between theory and observation appears to be dubious, the principle of verification fails "either by ruling out most of science as unscientific, or by ruling out nothing" (Thagard, 1978). Similarly, no observation is sufficient to guarantee falsification, because a theory may always be retained by introducing modifications.

The second force is the one popularized by Thomas Kuhn (1962/1970) in *The Structure of Scientific Revolutions.* I call those who subscribe to the paradigmatic account of science and believe that science is a social activity subject to historical, sociological and psychological influences, the revolutionaries. Stretched to its natural limits, "revolutionism" leads to a consensus theory of truth, and some of the revolutionaries have clearly courted irrationalism (Feyerabend, 1975).

The new realists believe that there is a world out there that is independent of the observer. But, unlike the traditionalists, and in line with the revolutionaries, they point out that knowledge is a product of social and historical factors. Consequently, there is no preinterpreted "given" against which our hypotheses can be tested and to which our theories can correspond. Instead, "it is precisely the task of science to invent theories that aim to represent the world.

Thus, . . . the practices of the sciences generate their own *rational* criteria in terms of which theory is accepted or rejected. The crucial point is that it is possible for these criteria to be rational precisely because, on realist terms, there is a world that exists independently of cognizing experience. Since our theories are constitutive of the known world but *not* of the *world,* we may always be wrong, but *not* anything goes" (Manicas and Secord, 1983, p. 401).

There is a significant difference between natural sciences and human sciences. In the former, the objects of inquiry are structures and not events. They are not empirical, in the sense that they are not observations per se. In the human sciences, on the other hand, the transformations of the pre-given by the human agency become more central to inquiry than empirical invariances do. Social phenomena are *emergent* and the generative mechanisms have to be found in the human intentions and interactions that operate in open systems. As Bhaskar (1982) puts it: "(1) criteria for theory assessment and development in the human sciences cannot be predictive, and so must be exclusively *explanatory;* and (2) social phenomena in general must be seen as the product of a multiplicity of causes, so that social events will be 'conjunctures' and social things (metaphysically) 'compounds'" (p. 278). Social structures and mechanisms that influence the activity of human agencies are themselves the products of human actions. We find nothing similar to this in natural sciences. When we move from normal psychosocial phenomena to psychical phenomena, we move into new areas that involve structures of far greater complexity and openness. Inasmuch as human intentions apparently have the ability to influence external physical systems as well as the intentions of other agencies, the problem is further compounded. Psychical structures are the most complex and open ones in the universe. Human beings do not simply operate in closed systems. Therefore, it would be naive to suppose that we can set up simple experimental conditions and obtain a relative closure and duplication of conditions that would ensure a replication in the sense of reproducing the results on demand.

A prominent physicist told me that in physical science, a discrepant finding, i.e., a result that conflicts with an accepted theory, has little chance of being noticed unless it is significant beyond the 10^{-7} level. Contrast this with the significance levels in behavioral sciences, where our statistical tables do not generally go beyond the .001 level. What I wish to emphasize here is that if replication on demand is unreasonable in behavioral sciences in general, it is clearly inappropriate for parapsychology. I do not believe that statistical replication, which

is less than producing a phenomenon on demand, is simply imperfect replication and that real phenomena are *in principle* repeatable in an absolute sense, if we have a perfect understanding of the crucial variables, i.e., if we could solve "the third-variable problem." On the contrary, I hold that such perfect understanding is unattainable in some areas, and parapsychology is one such area.

I have argued thus far that (a) replication is a nonessential aspect in normal science, (b) that it is an inappropriate criterion for distinguishing the genuine from the spurious in science and (c) that it is not an all-or-none phenomenon, but one that admits of various degrees depending on the relative closure that obtains in a given situation. I have also suggested that the instability and the elusiveness involved in psi phenomena are a function of the relative openness of the structures involved. Having said this, I am not about to say that the problem of repeatability is unimportant or irrelevant to parapsychology. Rather, I happen to think that, at this juncture, the problem of replication is perhaps the most important and crucial one and that the future direction of parapsychology depends on the resolution of this problem.

Replication in Parapsychology

I have three reasons for my emphasis on the repeatability problem in parapsychology. First, much of serious research in parapsychology is laboratory based. Parapsychology as a laboratory science presupposes that psi phenomena are in principle replicable. In other words, we cannot have a laboratory science without reproducible phenomena. Second, the rate of replication is a fair index of the frequency of occurrence of a given phenomenon. A knowledge of the frequency with which one could obtain psi in the laboratory is helpful in making an intelligent choice of a career in parapsychology. Third, applying psi for pragmatic use depends largely on our success in obtaining reliable results.

Frankly, a reading of the reviews of literature in the field makes me optimistic about replications of psi effects, more so than did my intuitive expectations of it based on the understanding of psi. For example, John Palmer's (1971) review of sheep-goat studies reveals that, in 13 of the 17 experiments that used standard methods of analysis, the sheep obtained higher scores than the goats did, with 6 of the 13 achieving statistical significance. Charles Honorton's (1977, 1978) reviews of micro-PK experiments with Schmidt's type of random event generators and experiments involving internal attention

states suggest a replication rate of approximately 50 percent. A review by Haraldsson (1978) of clairvoyance experiments in relation to the Defense Mechanism Test reveals a very high percentage of success. Of the 19 reported series of experiments involving ESP scores and EEG alpha activity, 15 contained significant effects of one sort or another. Of the 15, 9 significant studies had at least one effect that might be reasonably assumed to have been predicted (Rao and Feola, 1979). Carl Sargent's (1981) review of the English literature on the association between ESP and extraversion suggests that significant confirmations of a positive relationship occur at over six times the chance error.

My own count of 143 experiments that involved differential situations for the subjects shows that 95 (66 percent) of them registered differential scoring, when one expects this to occur by chance only in 50 percent of the cases. Even more striking is the large number of significant differences in the scores between the two conditions. In 72 of the series, the scoring rate between the two conditions is significantly different at or beyond the .05 level. This list does not include the experiments in which experimental and control conditions, which also provide a differential situation, are compared. If we had included these, the results would have been even more impressive. For example, if we consider the experiments that attempted to investigate the effect of hypnosis and meditation employing a within-subject design, we find differential scoring in 17 out of 20 studies. Of the 17, 14 are statistically significant. It is also interesting to note that when there are more than two conditions in an experimental situation, differential scoring does not seem to occur.

Methodological Problems of Replication

If the distinction I have made between *absolute* and *statistical* replication (Rao, 1981) is valid and the only kind of a replication we can hope for in psi phenomena is statistical replication, then we need to pay greater attention to the methodologic problems involved in such replication. Even behavioral scientists, who are typically unable to reproduce effects on demand, appear to pay scant attention to this question. In this connection the discussion of the general neglect of adequately dealing with Type II errors in behavioral sciences is quite illuminating.

One of the early papers on this subject is by Jacob Cohen (1962). He reviewed all the statistical analyses contained in Volume 61 of the *Journal of Abnormal and Social Psychology.* By using an ingenious

technique to detect the size of the effects, he found that the average power to detect small effects was .18. For medium and large size effects it was .48 and .83, respectively. Cohen's results indicate that the statistical power of the studies he reviewed is so ridiculously low that "the investigators . . . had, on the average, a relatively (or even absolutely) poor chance of rejecting their major null hypotheses, unless the effect they sought was large" (p. 151). Statistical power is extremely important in designing replication studies, because studies with low statistical power render us more susceptible to rejecting a valid hypothesis (Cohen, 1977).

In an interesting study, Amos Tversky and Daniel Kahneman (1971) reported the responses of audiences attending a mathematical psychology meeting and a session of the American Psychological Association convention to questions concerning replication. These responses provided considerable evidence that their typical respondent is a believer in what they call the law of small numbers. The believer in the law of small numbers "gambles his research hypotheses on small samples without realizing that the odds against him are unreasonably high." Further, "in evaluating replications, his or others', he has unreasonably high expectations about the repeatability of significant results" (p. 109). This is so because people generally have a tendency to view a randomly drawn sample as highly representative of the population, which is an unreasonable assumption when dealing with small numbers because it ignores sampling variations. Consider, for example, that you have reason to expect a correlation of .35 between extraversion scores and ESP scores. You would require an N of 33 to render $r = .35$ significant at the .05 level. Because of possible sampling variability, the probability of your obtaining a significant result ($r = .35$) is however only .50. Therefore if you design your experiment with a sample size of 35, your chances of accepting the null hypothesis are the same as rejecting it. In parapsychology, the belief in small numbers appears to be quite prevalent among researchers, as well as among their critics. Typically, parapsychological effects are very small, smaller than small size effects in psychology. When dealing with phenomena of very low signal-to-noise ratio, it is unreasonable to expect a high percentage of replication. This can be readily seen by computing statistical power in our experimental designs.

Let me illustrate this point with an example. Gertrude Schmeidler's group series of clairvoyance tests show that the ESP scores of sheep are significantly greater than those of the goats. Sheep obtained an average of 5.10 hits per run while goats scored at an average of

4.93. If we consider the mean difference between the sheep and the goats, for a standard run of 25 trials as the size of the effect, it is .17 in Schmeidler's original group series. Even though the magnitude of the effect is very small, the result is still significant because of the large number of subjects. She had in her group series 692 sheep and 510 goats (Schmeidler and McConnell, 1958).

It is of interest to note from a table provided by Palmer (1971), that of the 18 series of group sheep-goat experiments carried out after Schmeidler's study and using standard methods and analyses, only two gave statistically significant results. However, in eight other series the size of the sheep-goat effect was .17 or greater. A revealing example is one series by Eilbert and Schmeidler (1950) in which the size of the sheep-goat difference is .45, which is 2.65 times greater than the one in the original studies and is yet statistically nonsignificant because the sample consisted of only 19 subjects. The study of Eilbert and Schmeidler is clearly inappropriate as an attempt at replicating the sheep-goat effect, because its statistical power is so low and the odds are set heavily against rejecting the null hypothesis. Many of us sinned at one time or another by believing implicitly or explicitly in the law of small numbers and to some degree the problems of replication we confront in psi research may be attributed to this sin. Tversky and Kahneman (1971) recommend that "unrealistic expectations concerning the repeatability of significance levels may be corrected if the distinction between size and significance is clarified" (p. 110)—a distinction that is obscured by the emphasis on significance levels. While significance of a result depends critically on sample size, the size of an effect may be expected to remain the same regardless of sample size. This may indeed be the case in most areas of psychology. But in parapsychology, the problem is complicated by the generally observed decline effect.

I need hardly labor to emphasize to this audience the importance and the frequency of decline effects in psi research. From the time of Charles Richet (1923), George Estabrooks (1927) and I. Jephson (1929) to the present, declines in psi performance are observed. Such declines have sometimes provided strong internal evidence for psi. If in psi research we require lengthy series with large N's to provide necessary statistical power because of the very small size of the effects expected and if such lengthy series run the risk of inviting declines, we really face a serious internal contradiction, a contradiction between statistical power on the one hand and the nature of psi functioning on the other. We may need to strike a fine and delicate balance between the two and avoid a possible canceling effect. I am persuaded

that reviews of past research as well as planned future research are required to come up with an adequate recipe for an optimal design. The concept of psi quotient (Schmidt, 1970) deserves to be taken more seriously as a measure of the size of a psi effect and to be used as a guide in our design of replicatory experiments with due regard for statistical power. Also, we need to have greater insights into the nature of declines and the conditions that facilitate and those that work against their occurrence.

Despite the limitations of our methodology and the adverse effects of decline on our attempts to replicate, I find the frequency of replication indicated by the reviews of literature extremely assuring. With improved methodology and greater understanding, psi may indeed be found to be not so elusive an effect after all.

Predictive and Retrodictive Replications

Having said this, I must also mention a lurking fear which led me to a distinction I made between *predictive* and postdictive or *retrodictive* statements of replication (Rao, 1981). I believe we have in parapsychology an excellent case for retrodictive replications. Normally, judgments based on retrospective assessment of all the attempted replications of our effect would have predictive validity. But in parapsychology this may or may not be the case, because the act of making a prediction may itself influence the outcome. Psi may operate in such self-obscuring ways as to render retrodictive analysis devoid of any predictive validity. Therefore, the question of whether psi phenomena are in principle replicable, in the predictive sense, must be settled. With this in view, I have undertaken collaborative research with some of my colleagues in the field. We are carrying out 10 planned series of Ganzfeld ESP studies with 10 different experimenters under as secure and favorable conditions as we could possibly obtain, with the expectation that a good percentage of them would yield significant results.

This one attempt will not, of course, settle the question of replication. Rather I believe this study may provide a stimulus to do more research of that sort until we have adequate data to provide an answer. If the final answer is "yes," it would have enormous consequences for applying psi for pragmatic use. If, on the contrary, it is proven that psi is so self-obscuring that it will forever remain elusive and unpredictable, such a finding would have important consequences for the future course of parapsychology and its research methods and strategies. It would then in all probability cease to be a

laboratory discipline and become more an *experiential* subject. In either case, the study of psi would be legitimate and scientifically proper, but its significance to the human condition would be greatly different. Also, the kind of persons who would be attracted to psi research would be different, depending on which side the evidence leans.

At this point some of you may be wondering what form parapsychology as an experiential science would take. I cannot quite provide an adequate answer, as my interests in parapsychology are primarily laboratory-oriented, except to quote Rhea White who seems to come pretty close to what I have in mind. Writing on the future of parapsychology, she predicts: "We will pioneer a new type of research that will emphasize developing groups whose goal will be to attain consensual validation of inner experience as it relates to psi. Initially, it seems to me, each group should be composed of at least a few mystics, simply because they would already be familiar with some of the techniques the group would be developing; they would already have landed their anchors somewhere in the inner world or else would be consciously intending to do so; they would be practical in the sense of being interested only in what works; they would appreciate the importance of firsthand experience; and to an extent—as is often pointed out—they would speak the same language. Therefore, they would have an important role to play as members of a group whose goal would be to find commonalities of psi-relevant experience. In addition to identifying state variables relevant to the psi process that they already held in common, the group would try to enlarge its base in two ways: outwardly by adding more people, and inwardly by finding more psi-relevant commonalities" (White, 1983, p. 221, 222).

Conclusion

In summary, then, I have argued that the replicability issue is irrelevant for settling the questions of legitimacy of a scientific field or the genuineness of its findings. The problem of replicability is nevertheless important for parapsychologists, because the pursuit of psi in the laboratory and the possibility of applying its results for practical use presuppose that psi phenomena are, in principle, replicable. A distinction is made between *absolute* replication, i.e., reproducibility of results on demand, and *statistical* replication, i.e., repeated but nonuniform observation of an effect or a phenomenon that is not coincidental. Absolute replication is an absurdity when dealing with complex structures and emergent events that operate at different

levels within relatively open systems. Low-size effects masked in high-level noise are replicable only in a statistical sense.

Statistical replication is of two kinds—retrodictive and predictive. There is already evidence that psi phenomena are replicable in the retrodictive sense. Research is needed to ascertain whether psi results are replicable in a predictive sense.

I have pointed out that statistical power in psi experiments is generally quite low and that it may explain, in part, some of the failures to replicate. Statistical power may be increased by increasing sample size, but the frequent occurrence of decline effects in psi research creates special problems we must deal with. I have suggested that we make more frequent and better use of the concept of psi quotient or a variant of it and attempt to strike a workable balance between the size of the effect and the length of the series to achieve better replication in psi experiments.

BIBLIOGRAPHY

Bhaskar, R. *A Realist Theory of Science.* Leeds, England: Leeds Books, 1975.

Bhaskar, R. "Emergence, explanation and emancipation." In P. F. Secord (Ed.), *Explaining Social Behavior: Consciousness, Behavior, and Social Structure.* Beverly Hills, CA: Sage, 1982.

Carnap, R. "The methodological character of theoretical concepts." In *Minnesota Studies in the Philosophy of Science.* (Vol. 1). Minneapolis: University of Minnesota Press, 1956.

Cohen, J. "The statistical power of abnormal-social psychological research." *Journal of Abnormal and Social Psychology,* 1962, *65,* 145–153.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences.* (Rev. ed.) New York: Academic Press, 1977.

Collins, H. M. "Science and the rule of replicability: A sociological study of scientific method." Paper presented at the Annual Meeting of the American Association for the Advancement of Science, Washington, D.C., February 17, 1978.

Eilbert, L. and Schmeidler, G. R. "A study of certain psychological factors in relation to ESP performance." *Journal of Parapsychology,* 1950, *14,* 53–74.

Estabrooks, G. H. "A contribution to experimental telepathy." *Bulletin V.* Boston: Boston Society for Psychical Research, 1927.

Feyerabend, P. R. *Against Method: Outline of an Anarchistic Theory of Knowledge.* London: New Left Books, 1975.

Hanson, N. R. *Patterns of Discovery.* New York: Cambridge University Press, 1958.

Haraldsson, E. "ESP and the defense mechanism test (DMT): A further validation." *European Journal of Parapsychology,* 1978, *2,* 104–114.

Harré, R. *The Principles of Scientific Thinking.* Chicago: University of Chicago Press, 1970.

Honorton, C. "Psi and internal attention states." In B. B. Wolman (Ed.), *Handbook of Parapsychology.* New York: Van Nostrand Reinhold, 1977.

Honorton, C. "Replicability, experimenter influence, and parapsychology: An empirical context for the study of mind." Paper presented at the Annual Meeting of the American Association for the Advancement of Science, Washington, D.C., February, 17, 1978.

Jephson, I. "Evidence for clairvoyance in card-guessing." *Proceedings of the Society for Psychical Research,* 1929, *38,* 223–268.
Kuhn, T. S. *The Structure of Scientific Revolutions.* (2nd ed.) Chicago: University of Chicago Press, 1970. (1st ed. 1962).
Manicas, P. T. and Secord, P. F. "Implications for psychology of the new philosophy of science." *American Psychologist,* 1983, *38,* 399–413.
Moss, S. and Butler, D. C. "The scientific credibility of ESP." *Perceptual and Motor Skills,* 1978, *46,* 1063–1079.
Palmer, J. "Scoring in ESP tests as a function of belief in ESP. Part 1. The sheep-goat effect." *Journal of the American Society for Psychical Research,* 1971, *65,* 373–408.
Polanyi, M. *Personal Knowledge.* London: Routledge and Kegan Paul, 1950.
Polanyi, M. *The Tacit Dimension.* London: Routledge and Kegan Paul, 1967.
Rao, K. R. "On 'The scientific credibility of ESP.'" *Perceptual and Motor Skills,* 1979, *49,* 415–429.
Rao, K. R. "On the question of replication." *Journal of Parapsychology,* 1981, *45,* 311–320.
Rao, K. R. and Feola, J. "Electrical activity of the brain and ESP: An exploratory study of alpha rhythm and ESP scoring." *Journal of Indian Psychology,* 1979, *2,* 118–133.
Richet, C. *Thirty Years of Psychical Research.* New York: McMillan, 1923.
Rosenthal, R. *Experimenter Effects in Behavioral Research.* New York: Irvington Publishers, 1976.
Sargent, C. L. "Extraversion and performance in 'extra-sensory perception' tasks." *Personality and Individual Differences,* 1981, *2,* 137–143.
Schmeidler, G. R. and McConnell, R. A. *ESP and Personality Patterns,* New Haven, CT: Yale University Press, 1958.
Schmidt, H. "The psi quotient (PQ): An efficiency measure for psi tests." *Journal of Parapsychology,* 1970, *34,* 210–214.
Thagard, P. R. "Why astrology is a pseudoscience." *Proceedings of Philosophy of Science Association,* 1978, *1,* 223–224.
Tversky, A. and Kahneman, D. "Belief in the law of small numbers." *Psychological Bulletin,* 1971, *76,* 105–110.
White, R. "The future of parapsychology." *The Journal of Religion and Psychical Research,* 1983, *6,* 220–226.

DISCUSSION

WALKER: I really couldn't agree more with most of what Dr. Rao had to say. I want to add a few exclamation points to some of what he said. I have on a number of occasions tried to get across the idea that Dr. Rao called the law of small numbers. I would like to point out that even when the phenomena are assumed to be real, in many of the efforts to replicate one should not expect to find that the phenomena do appear for exactly the reason that Dr. Rao stated.

My second point has to do with the matter of statistical power, of being able to discriminate the existence of an effect, to find the presence of a small effect. Dr. Rao gives the number of .18 as the size of an effect that can just barely be discriminated in most experiments. But there are very good reasons for believing that we

are looking at an effect that is several orders of magnitude smaller than that. This, perhaps, is a point that should be stressed in talking to many of our critics.

My third point is that I would tend probably to disagree with the number of 10^{-7} for the significance, as you were told by a physical scientist. I really don't remember the exact numbers for it, but in the original Eddington experiment to look for the deflection of light passing the sun as a confirmation of Einstein's theory of general relativity, I think if you check the numbers there you will find it was nothing like that and yet the results were met with overwhelming acclaim that confirmation had been found. I think you will find actually that the numbers that Eddington came up with deviated by about 20 percent from the deflection that was expected, which was twice as great as the one Newton predicted. And, therefore, it couldn't have been a 10^{-7} level of significance. This is probably a number that refers to a lot of the elementary particle experimental work that goes on now. This goes back, I think, to this question of a paradigm at the time of the test of Einstein's theory. There was a much greater expectation that, indeed, this would be found to be true. When there are so many elementary particle theories floating around now, one must come up with a very highly significant result in order to pin down the reality of one effect vis á vis a prediction of something that happened. But, I want to point out that number 10^{-7} because I will be coming back to something of that sort when I give my talk.

RAO: In regard to your third point, my authority is a professor of physics at the University of Virginia, who gave a paper not too long ago, at a meeting of the Society for Scientific Exploration. This is the essential point he made. Maybe you are right, you know more than I do about physics.

WALKER: Let me just add that physicists love to beat their chests about how great their measurements are. There may have been more of that pride among physicists in what he was stating than there was reality that could be borne out by the history of physics.

STANFORD: Dr. Rao, I also had an exclamation point with regard to our seeming belief in the law of small numbers. I have been concerned about this problem for years. In your paper, at least by implication, you asked whether decline effects might be in some way modified, the way we would attempt to deal with the law of small numbers. Normally we would expand the number of observations in our experiments. I think it is important to keep in mind that as we attempt to extend the size of the experiment, say, add to 40 subjects

40 more subjects, making 80, we may lose something in the process. I think that might be one of the reasons why we sometimes observe what seems like an adherence to the law of small numbers. I wish it were the only reason. I think we are not aware enough, sometimes, of the statistical power considerations. What I want to suggest is that there may be ways around that kind of problem, such as bringing in a variety of experimenters so that the experimenters don't become bored with testing subjects and an experimental session doesn't become something at the end of an experiment that it wasn't when it began. This gets into some methodological considerations that we won't be able to explore at length here.

I perceive what, in my mind at least, is an implicit contradiction in your presentation. Maybe you can clarify it for us. You certainly wound up talking about the importance and need of replicability, at least in this field. I fully acknowledge this importance if we are going to try to do laboratory science, because if we are going to do experiments we have to set up conditions and expect them to cause something perhaps to happen. But in the beginning of your talk you seem to strongly downplay the importance of replicability. First of all, how do scientists ever come to agree upon new knowlege? An assumption that nature is stable and that if we make the same kinds of observations again we will be able successfully to make those same observations, is precisely what protects science from the kind of subjectivity that you find in, say, religion. Some recent philosophers who have stressed a sociological perspective have said that whether we are going to get many attempts at replication or not is going to depend on how skeptical people are about the original findings. It is a *sine qua non* of science that we assume that nature at least is not going to change its rules. And that if for any reason we decide to go back and check, we can. I think there have been examples in the history of science where we have not checked enough initially and twenty or thirty years later we had to come back and check and find out that we had to change our opinions.

RAO: With regard to the first point, I quite agree with you that the decline phenomenon is a serious problem. In fact, what I am suggesting is that we pay more attention to it, do research into the decline question, so that we can moderate declines and therefore maximize the size of the sample. I think there we are in essential agreement.

With regard to your second and more provocative remark, I must add that I personally don't see any contradiction between the first part and second part of my paper. In the first part I am arguing that

replication is not an essential part of science. And, therefore, even if parapsychological phenomena are not replicable, we can have a science of parapsychology and I have given the reasons why that is the case. In the second part, I have argued that, for us in parapsychology at this juncture, adopting the kind of strategies that we do in the field, replication is important. I also emphasized that we seem to be not too far away from the point where we can feel fairly comfortable with the situation as we wish it to be.

Now, what you perceive to be a contradiction between the two is resolved if you look at the kind of philosophy of science I agree with. That is the new realist philosophy of science that holds there is a reality out there, but that is not the reality with which you are doing your research. You are doing your research with the reality that you have in your mind, what you perceive to be the case. As long as you are dealing with open structures, with hierarchical organizations, there is no way you can have an exact duplication of the conditions so that you can get exactly the duplication of the result. In my judgment it is unreasonable to expect 100 percent replication in parapsychology. As a matter of fact, such is not the case in other branches of behavioral science. In other words, you are saying in a positivistic sense that replication means that it is (a) possible to duplicate conditions and (b) when you duplicate those conditions you get exactly the same result. I am questioning both of those points of view. And when you do that, what I say logically follows.

STANFORD: I think it is unreasonable, in many areas of behavioral science, to expect to get this kind of 100 percent replicability precisely as you say. We do not know whether or not we have established the same kinds of conditions and should observe the same kind of consequences. But what I am still wondering is whether that in any way invalidates the importance of whether it is not an ideal to move in that direction, even though it is admittedly very difficult.

RAO: Well, I'm not talking here about an ideal for science, certainly not an ideal for parapsychology. I'm talking about the possibility of progress that we can make by employing a particular stance or a strategy. In that sense it seems to me it is completely irrelevant whether or not there would be the kind of replication that you are talking about. Replication is important because I am interested in making use of the phenomena I am studying. Replication is important because, as John pointed out earlier, it is what brings more funds into the field and better minds to work with and rapid progress in the field as well. And replication is important also because, as a laboratory scientist, I need to know how frequently I can observe

the phenomenon, which obviously is necessary to study it. I want to be sure before I go into the field what is the prospect and the possibility of dealing with these phenomena. How frequently can I encounter them? And replication, I think, is a way of measuring the possible frequency of encountering the phenomena in the laboratory situation. So for these pragmatic and not, unfortunately, idealistic reasons, I think replication is important. But if you are talking about it in principle, for idealistic reasons, I still argue that I don't care whether the phenomena replicate themselves or not, because that is not what science is about.

HONORTON: I just want to reinforce what I think is an extremely important point and that is that we pay more attention to effect size and not confuse it with statistical significance. The Pearce-Pratt experiment had a P value that was astronomically significant. It was 10^{-22}. But the effect size was very small. It was an average of maybe two hits per run above what would be expected by chance. If you do a random number generator PK experiment with a million trials you can have a less than a tenth of one percent deviation from chance to produce a highly significant result. Now that is a meaningful result, but the fact that that might be associated with a P value of 1 in a million doesn't make it a strong effect. And I think there is another aspect to effect size that has to be taken into consideration. We are talking about replication and we are using a significance level as a measurement of replication. But what we really ought to be using, it seems to me, is the confidence interval. Whenever we take a statistical measurement, all we are doing is sampling from an underlying distribution and each experiment represents one point estimate of the likely mean of that distribution. If we do a Ganzfeld experiment, for example, that has a scoring rate that is twice chance, 50 percent instead of 25 percent with say 30 or 40 trials, the confidence interval there is something like 15 percent, which means that the true effect mean could be anywhere from slightly above chance on up to maybe 70 percent or so. So, when we are attempting to evaluate replication, it is very important that we be able to see quantitative data from the studies that say more than what the P value was, whether it was significant or not. And many of the non-significant failures to replicate in all areas of parapsychology very often do not give more than the statement that the results were not statistically significant. For all we know the results could be within the confidence interval of a genuine real effect. I also think that we need to develop some fairly sharp criteria of what constitutes an acceptable level of replicability. Dr. Beloff mentioned that his 50

percent criterion of ten years ago was arbitrary. Well, the 5 percent criterion of statistical significance is also quite arbitrary. But I think we need some criterion like that to evaluate replicability. First, so that we can keep John from changing his mind every ten years when the field begins to approach his earlier criterion and because we really need to have some clear cut criterion for evaluating what we mean by replicability. Marilyn Schlitz was saying earlier that she doesn't think we could agree on a definition of psi. I hope she is wrong about that, because otherwise we might as well go to lunch and then back to where we came from!

RAO: I agree entirely with your first point. I think some of the work you are doing to assess the significance of replicability is very important and we look forward to hearing from you on that later on in the program.

In regard to the second point, I think, if you are talking about a logical criterion, we already have that in the statistical criterion—that is that replication is not simply coincidental. If you do 100 studies, five of them are going to be significant anyway, just by chance. But if the replication rate is statistically significant, then, of course, whatever the criteria are that you want to use for the statistical significance—5 percent or 1 percent—you have a logical criterion for evaluating a phenomenon or an experimental procedure that has been validly replicated or not, but that is not, I think, what we are interested in very much. That is good enough to deal with the critics, but from our own point of view we need to have a workable criterion and that is more pragmatic. And this pragmatic criterion is dictated by what use we want to make of these results. If we are going to make use of the results to make applications now, what are the procedures? Do we have to increase the signal to noise ratio? What magnitude of effect do we need to obtain the kind of results which make it possible for us to apply the phenomena? In that case, I think that it has to be some kind of pragmatic determination which is dependent upon our own needs as researchers to obtain phenomena—as researchers who want to apply the phenomenon to possible practical use. This is bound to be changing as our research interests change. So I don't particularly find fault with John for changing his criteria. In a sense, I feel good about it because, I think, while he said he is less optimistic I feel he is more optimistic; therefore he is raising his criteria upward not going downward.

BLACKMORE: In your discussion of statistical power you seem to assume that there is a real effect there and that if you have more statistical power you are more likely to get significant results. This

obviously leads to a very specific prediction which can be tested. That is, with more statistical power and larger sample size you should more often get significant results. This has been tested several times in the past and most recently by Ray Hyman and Chuck Honorton, who have been discussing measures with Ganzfeld work. Unfortunately, the results are not clear cut because the two of them seem to have come to different conclusions. My reading of that evidence is that there isn't the kind of clear, clean effect that we would hope for if we, indeed, have a real effect that we have been missing because we haven't had enough statistical power. Now, that doesn't necessarily mean there is no real effect there. Rex Stanford may be quite right in suggesting that longer experiments are in some essential way different from shorter experiments. The later trials become boring. But, if that is the case, then we can't use the sort of ostensibly rational arguments that you are using about statistical power. In fact, we are in even deeper trouble than you seem to imply.

RAO: Well, I must disagree with you on many statements that you have made. First of all, I don't imply that increasing the sample is necessarily going to result in obtaining significance in parapsychological research, for the reason I have mentioned, that there is a decline problem in parapsychology. So what I want to focus on is to pay more attention to declines and study them more carefully, so we can determine what is the optimum size of the sample that you should have in a given set of experiments. If you are dealing with individual testing, or with group testing, if you are dealing with the Ganzfeld type of free-response situations or with guessing cards or a random number generator you have different variables and you have to carefully consider for a given experimental situation what is the optimum length that you can have without occurrence of the decline effect. And so my purpose in calling attention to this is to interest research workers in taking the problem more seriously and determining the limits that we can reasonably set to the decline effect. I am not aware if there have been any review studies made in the field to determine exactly whether this issue has been settled statistically or otherwise.

BLACKMORE: You are quite right to emphasize the importance of declines. Now it may be that there is an optimum before the decline or the other effects get too bad and while the statistical power is building up. But what if we should find that one has begun to cancel out any effect that we can find before the others built up? What do we conclude then?

RAO: Well, I can't really answer the hypothetical question. The results of research will have to speak for themselves. But what I am saying is that it is the way that you have to proceed in your research. Given the results that we have on hand the situation seems to be more promising than depressing.

HONORTON: In relation to the point Sue Blackmore mentioned concerning the Ganzfeld studies, in my evaluation of the Ganzfeld research, if you look at the studies that are optimized for success in the sense of having longer Ganzfeld duration, using some kind of relaxation procedure, having experienced subjects, giving subjects the option of selecting their own sender, these studies tend to be on the average about 47 sessions per study compared to the overall average for the Ganzfeld of around 37 sessions. And although this difference is not statistically significant with these relatively small numbers of studies involved, a high proportion of the significant Ganzfeld studies fall in that one group. So that the point that I would make is that in considering the relationship between the effect size and the number of trials, you also consider the other factors such as conditions that would seem to be optimizing of the effect if the effect is a real one.