of a discipline it turns to meta-analysis to answer research questions or to resolve controversy (e.g., Greenhouse et al., 1990).

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

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

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

ACKNOWLEDGMENT

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

Comment

Ray Hyman

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

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

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

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

the reported data at face value and focuses on the statistical interpretation of these data.

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

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

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

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

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

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

Utts mistakenly asserts that my subcommittee on parapsychology commissioned Harris and Rosenthal to evaluate parapsychology experiments for

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

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

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

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

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

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

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

Utts' discussion of the ganzfeld debate, as I have indicated, also shows unfamiliarity with details. She cites my factor analysis and Saunders' critique as if these somehow jeopardized the conclusions I drew. Again, the matter is too complex to discuss adequately in this forum. The "factor analysis" she is talking about is discussed in a few pages of my critique. I introduced it as a convenient way to summarize my conclusions, *none of which depended on this analysis*. I agree with what Saunders has to say about the limitations of factor analysis in this context. Unfortunately, Saunders bases his criticism on wrong assumptions about what I did and why I did it. His dismissal of the results as "meaningless" is based on mistaken algebra. I included as dummy variables five experimenters in the factor analysis. Because an experimenter can only appear on one variable, this necessarily forces the average intercorrelation among the experimenter variables to be negative. Saunders falsely asserts that this negative correlation must be -1. If he were correct, this would make the results meaningless. But he could be correct only if there were just two investigators and that each one accounted for 50% of the experiments. In my case, as I made sure to check ahead of time, the use of five

experimenters, each of whom contributed only a few studies to the data base, produced a mildly negative intercorrelation of −0.147. To make sure even that small correlation did not distort the results, I did the factor analysis with and without the dummy variables. The same factors were obtained in both cases.

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

CONCLUSIONS

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

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

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

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

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

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