Comment

Robert L. Morris

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

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

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

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

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

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

Robert L. Morris occupies the Koestler Chair of Parapsychology in the Department of Psychology at the University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, United Kingdom.

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

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

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

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

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

misses estimated; perhaps Cohen's h greatly underestimates effect size when very low probability events (less than 1 in 50 for heart attack in the placebo condition and less than 1 in a 100 for aspirin) are involved. I'm not a statistician and thus don't know if there is a relevant literature on this point.

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

Comment

Frederick Mosteller

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

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

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

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

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

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