More on Methodological Issues in Free-Response Psi Experiments

J. E. KENNEDY

(Original publication and copyright: *Journal of the American Society for Psychical Research*, 1979, Volume 73, pp. 395-401)

INTRODUCTION

Since the publication of my previous paper on free-resppnse methodology (Kennedy, 1979), several developments have taken place that deserve comment. In his paper in this issue of the *Journal*, Charles Honorton has raised various issues that merit further discussion and, in addition, some matters that should have been included in my previous paper have recently come to my attention. I will first comment on some of the issues raised by Honorton, and then discuss the other topics.[1]

ON THE PAPER BY HONORTON

The purpose of my previous paper was to point out some methodological problems that should be avoided in future free-response work and to indicate studies that had been reported with these problems. I did not argue for or against the hypotheses that were being investigated in these studies other than to point out that the methodology made particular results questionable or exaggerated. I hoped that, in addition to preventing these problems from arising in the future, the paper would stimulate those involved in the research to present further information and analyses that would clarify the questionable results of the published studies. Honorton has provided some new information and analyses in an effort to answer various questions that I raised. At the same time, he has made numerous comments and criticisms that need some discussion.

Honorton's main criticisms center around the way that the subject matter of my paper was delimited and presented. For example, he thinks that some of the methodological problems discussed

---

[1] For further discussion of these topics, see the "Correspondence Section" in this issue of the *Journal.—Ed.*

were highly selective and/or presented without an empirical assessment of their actual effect in either individual studies or lines of research as a whole. A large part of his paper is devoted to discussing these points. However, as noted above, in many cases I did not possess the information required to accurately assess the extent to which the potential problems actually affected the results of particular experiments. I also specifically avoided (a) attempting to evaluate the overall impact of the methodological problems on the different lines of research, and (b) any analyses that required using all the studies in particular lines of research. The most appropriate way to carry out analyses and draw conclusions involving entire lines of research would be to make a table for each line of research in which all the experiments would be listed along with the presence or absence of various methodological problems (e.g., improper statistics, number of analyses, possibility of sensory cues, etc.). In order to draw firm overall conclusions, all the methodological factors must be considered simultaneously, not each in isolation and ignoring the others. I think it was perfectly proper to limit the scope of the paper to exclude this major undertaking. At the same time, I also chose not to offer either informal impressions on the matter or incomplete analyses.

In the same vein, Honorton thinks that the coverage of the paper should have been expanded to include methodological problems in forced-choice experiments and in several places he offers informal impressions as to the extent of such problems. The original stimulus for writing the methodology paper came when, during the course of reviewing the free-response literature for another project, the recent occurrence of various dubious or erroneous practices led me to believe there was a need for such a paper. The scope of the paper was limited to free-response experiments because they were of most concern to me at the time and because the various kinds of problems typically take somewhat different forms in these experiments than in forced-choice experiments. For example, the primary (and infrequent) abuse of the CR statistic in free-response experiments has to do with its application to small sample sizes. The "pervasive abuse of the CR measure in forced-choice experiments" that Honorton refers to (p. 386) involves a different problem which has to do with the generalizability of results in process-oriented research (Stanford and Palmer, 1972). A discussion of this topic would have been included if I had noticed any free-response studies with the problem (very few process-oriented free-response experiments have been carried out). Since spending more time justifying the purposes and range of the previous paper would serve little purpose, I will go on to comment on some other issues that Honorton discusses.

Honorton comments (p. 397) that I unfairly implied that a study by Terry and Honorton (1976) was biased due to data selection. In fact, I stated that the discarded data should be reported and explained why this is so without making any implications about the outcome of that report. The hypothesis of data selection was and *will remain* a viable alternative to the ESP hypothesis until it is shown (empirically) that the data are not in accordance with the selection hypothesis. While the absence of a decline in the selected data is favorable to the ESP hypothesis, the most important analysis is to show that the discarded data do not have a lower scoring rate than the selected data—particularly the first few trials of the selected data.

Further clarification of the topic of multiple analyses also seems needed. As implied in Honorton's paper, the nature of multiple analyses in free-response experiments is in general somewhat different from that which has historically occurred in card-type experiments. Thus, analyses for declines, displacement effects, etc., are looking for psi effects beyond the main effect of overall scoring, while most free-response experiments have exclusively evaluated the overall scoring rate. It is well known that the problem of multiple analyses arises when several different effects are being considered. The point I wanted to make was that, although it is usually much less obvious in the reports, *the problem of multiple analyses also occurs when several analyses are done to evaluate one effect.* The examples of studies that found different results when different judges and/or statistical procedures were employed were given to indicate that diverse outcomes do sometimes occur in these situations and, therefore, selection of the most favorable analysis can be a problem.

The method Honorton used to estimate the number of unre-ported nonsignificant studies needed to reduce to chance the combined significance of all reported experiments, like the methods for calculating the combined significance itself, assumes that only one analysis was carried out per study. If essentially the same analyses were - carried out but correcting (somehow) for the number of analyses in each experiment, the figures for the combined results would be less significant than those that have been reported. This is not to say that they would be nonsignificant, but only that in the present form they are exaggerated to an unknown degree.

When evaluating entire lines of research, the multiple analyses question must be considered for experiments that are nonsignificant as well as for those that are significant. For example, in an experiment investigating ganzfeld and induced relaxation (Wood, Kirk, and Braud, 1977) three planned statistical measures were used to evaluate the overall evidence for psi and the relative effec-

tiveness of each of four separate conditions. The results for all analyses were nonsignificant. While it is not clear what results would have been required for the experiment to be considered significant, there would seem to be numerous possibilities. When evaluating lines of research the number of analyses may be as relevant as the number of experiments. However, the unknown degree of dependence between various analyses in any given case makes precise handling of these situations very difficult.

Making an evaluation of overall lines of research is an important matter in areas such as parapsychology where many relevant variables are difficult to control and thus, results are not reliable. One needs to draw conclusions from groups of experiments, but at the same time not try to be more precise than the data base can withstand. If all studies are included in these evaluations, then methodologically weak experiments may distort the conclusions. On the other hand, if certain studies are excluded, then the criticism of data selection can be raised. The best strategy may be to consider all studies and note the presence of all methodological questions, as was suggested for the tables recommended above. When considering the evidence for an effect, the total number of studies and overall success rate, evaluated in context of the methodological issues, may be the most valuable summary.

ADDITIONAL TOPICS

Greville's Method

When discussing the use of Greville's method in my previous paper, I noted (as had numerous authors before me) that the usual formula may lead to inflated results for small sample sizes, that the problem has not been carefully studied, and that computer programs could be written to properly handle small sample sizes. Such programs would examine and count all permutations of the existing data and find the corresponding exact probability. I have recently had some experience that leads me to believe that the inflated results in applying the formula may be more severe than has generally been realized. Therefore, until a more thorough investigation has been made, the computer counting of permutations is recommended for closed-deck free-response experiments (i.e., a closed pool of targets and corresponding responses) since the sample sizes are typically small.

Sensory Cues

Marks and Kammann (1978) have recently pointed out a possible sensory cues problem in a remote viewing experiment. I will discuss the problem in more general terms here. These concerns can

be raised for several experiments in the remote viewing and dream telepathy lines of research.

If a subject does several trials in a free-response experiment and gets feedback of the target after each trial, then the later responses could contain cues as to what the targets were on previous trials. A judge who sees all the responses may be able to utilize the cues. In a closed-deck judging situation, cues giving information as to what target was used for other trials also provides information on which targets are incorrect for a particular response, thus introducing biases in two ways. The problem still arises in the open deck case, but is less severe since information about other trials is of no help in judging a particular response. The extent to which such cues could have noticeable effects would depend on the type of cues, the sensitivity of the judge to the cues, and the details of the judging procedure (e.g., open versus closed deck and randomization of the targets).

Any gross sensory cues such as directly or indirectly mentioning previous targets could be edited out of the response transcripts. Of course, any editing must be carefully justified since it could also lead to biases. Other more subtle cues can easily be imagined that could not be identified and edited out. For example, a subject's responses could specifically avoid descriptions that would apply to the previous targets, thus introducing a bias in the responses that could have an effect in closed deck situations. Another example is that the first part of a subject's response could be based on recent experiences—including previous trials—while the later part could be due to imagination and/or ESP. A judge sensitive to this pattern could receive cues as to the targets for other trials. Numerous other equally speculative examples can be imagined.

At this point, it might be best to comment on the issue of the plausibility of hypotheses like these. Clearly, when there are dramatic correspondences between target and response, subtle sensory cue explanations are not applicable. However, typically hits which are not dramatic enter into the statistical analyses and it is with these cases that the subtle sensory cue hypotheses must be entertained. As Honorton notes, the sensory cue explanation is not particularly plausible in most cases. But, of course, some people find ESP an even more implausible explanation for significant results. Since research into subliminal perception and experimenter expectancy effects indicates that subtle cueing can sometimes have effects, it seems preferable from both a scientific and an economic point of view to use experimental designs that exclude these alternative hypotheses rather than to debate their plausibility in particular situations.

There may be situations—for example, when using remote

viewing targets—in which designs that completely preclude sensory cues may require impractical investments of time by the judges. In such cases, the less controlled procedures may be appropriate for exploratory work since the likelihood of sensory cues can be diminished by checking the transcripts. It should be kept in mind, however, that the same experimental designs could be used for investigating subtle sensory processes and the ultimate interpretation of the results may depend on each individual's opinion about the relative plausibility of ESP versus sensory cues. As further knowledge is acquired about the types of information used by successful ESP judges and about the limitations of sensory cues, the importance of personal opinion in interpreting these results will decrease. For the present, when at all possible, procedures that *completely* preclude sensory cues should be employed for psi experiments.

It is therefore recommended that closed-deck free-response procedures not be used if trial-by-trial feedback is given and the subject does more than one trial. In general, with closed-deck judging, the subject (and perhaps the personnel interacting with the subject) should be kept blind to all the other targets in the series as well as to the current target, and the target order should be randomized before being presented to the judge. In cases in which immediate feedback is given, an open-deck judging procedure using *unedited* transcripts is preferred. The simplest way to completely preclude the possibility that sensory cues in the transcripts could influence the judge is to have the responses judged one at a time in the same sequence they were produced by the subject. If the judge is prevented from changing his evaluations after seeing later responses, there can be no problems from any cues in later transcripts. An open-deck procedure with the subjects judging their own responses, as is commonly employed in ganzfeld experiments, is a straightforward (and successful) design that is free of the sensory cue problems discussed here.

REFERENCES

KENNEDY, J. E. Methodological problems in free-response ESP experiments. *Journal of the American Society for Psychical Research,* 1979, 73, 1-15.

MARKS, D., AND KAMMANN, R. Information transmission in remote viewing experiments. *Nature,* 1978, 274, 680-681.

STANFORD, R. G., AND PALMER, J. Some statistical considerations concerning process-oriented research in parapsychology. *Journal of the American Society for Psychical Research,* 1972, 66, 166-179.

TERRY, J. C., AND HONORTON, C. Psi information retrieval in the ganzfeld: Two confirmatory studies. *Journal of the American Society for Psychical Research,* 1976, 70, 207-217.

WOOD, R., KIRK, J., AND BRAUD, W. Free response GESP performance following ganzfeld stimulation vs. induced relaxation, with verbalized vs. nonverbalized mentation: A failure to replicate. *European Journal of Parapsychology*, 1977, 4, 80-93.

Institute for Parapsychology
College Station
Durham, North Carolina 27708

Other Methodology Articles