carry out a global significance test for such single hypotheses to which a superordinate null hypothesis can be assigned.

It should be clear that by performing global significance tests many psi experiments must lose their significance. I remember, though, that I also mentioned the interexperimental selection above, to whose avoidance, at the least, all similar psi experiments should be combined and submitted to a global significance test. Through such a "meta-analysis," on the other hand, the significance may increase so that the single experiment loses part of its meaning.

My second theme is the reduction of beta errors in the statistical evaluation of psi experiments. The problem is to increase the statistical efficiency (or power) of the significance tests in such a way that--despite the avoidance of selection errors--minimal psi effects can be statistically detected. I confine myself to two different questions, both of which are of considerable importance to the practice. The first question is: which are the statistically optimal methods for correcting a given selection or for combining single results which shall undergo a global significance test?

Here, it can first be answered that for any selection of a single result there is a simple statistical correction possible that replaces the global significance test. An approximate formula for this purpose requires that one multiplies the p value of the selected result with the number of given results. Naturally, in this manner, the p value will be strongly increased so that the statistical significance will in most cases disappear, as in the case of a global significance test. Nevertheless, this is a universal and very simple method of correcting intra- or interexperimental selection.

Most of the other methods consist in weighted combinations of the single results so as to attain a most efficient global significance test. In the case of standard psi experiments that seems trivial because one needs only to add the different hits, whose sum can be evaluated with a CR just as well as the separate results. However, an analysis of intra- and interindividual distributions of psi scores shows that the simple addition of hits is one of the statistically least efficient methods, even for the aggregation of small experimental units such as individual runs. The reason for this lies in the strong variability of psi scores, which can vary even in a bipolar fashion between psi-hitting and psi-missing so that the hit deviations cancel out each other. Therefore, I have suggested special (nonlinear) transformations weighting the single scores according to their size. Finally, following the method of the likelihood quotient, I came to a measure which is statistically most efficient for strongly varying psi scores and is a linear function of the well-known "run-score variance."

The second question refers to the identification of permissible

forms of selection which one could use to increase the statistical efficiency. For example, the above definition of selection error allows one to exclude any partial results from the global significance test of an experiment if the exclusion ensues according to a criterion that, under the null hypothesis, is independent of the respective results. If one, in this way, discovers certain clues that particular experimental situations, certain subjects, certain variables, etc., could be unsuccessful, one is allowed to eliminate them as is. This can be a great advantage because every nonsignificant partial result reduces the significance of the total result.

In the global statistical evaluation of a multivariate experiment, one should, further, reduce correlated criterion or predictor variables to a smaller number of factors by performing a factor analysis, because the statistical efficiency in the case of correlated variables decreases with the number of variables. Finally, the so-called extreme-group method should be mentioned, according to which one is allowed to eliminate the middle cases of the distribution of a variable when calculating correlations. For example, one could eliminate all the chance-scoring subjects of a correlational study, if enough psi-hitters and psi-missers remain. The correlations between psi variables and other variables could, in that way, become much more significant.

I am afraid my explanations will not lead to a decisive change in the statistical methods of parapsychologists. When I pointed to the problem of statistical selection errors at the 1980 PA Convention in Reykjavik, it also did not have any considerable effect. One must, apparently, turn to the psi skeptics to attain such effects. Probably, selection errors serve the general psychological tendency to synchronize the given empirical data with one's own expectations regarding reality. Therefore, the final demand can only be to answer one's own ways of acting with increased self-criticism, even in such an objective area as mathematical statistics. Otherwise, those cynics will be confirmed who always have contended that, with statistics, one can prove everything.

EVALUATING FREE-RESPONSE RATING DATA

Sybo A. Schouten† and Gert Camfferman (Parapsychology Laboratory, University of Utrecht, Sorbonnelaan 16, 3584CA Utrecht, The Netherlands)

During the recent decades the use of forced-choice methods in experimental research in parapsychology has gradually declined in favor of free-response techniques. A disadvantage of free-response techniques is that they are rather time consuming. The

discrepancy in time investment between free-response and forced-choice studies seems only acceptable if it can be proven that either free-response studies are more sensitive for detecting ESP or that knowledge is gained from the process analysis which free-response studies allow. These two potential advantages of free-response studies require, however, more sensitive techniques for analyzing free-response data than evaluations based on hit/miss ratios which are used with forced-choice methods.

An evaluation method often used in free-response studies is one that employs different target sets for each trial and has the subject assign ratings to all pictures of the set. A target set consists of a number of pictures from which one is randomly selected to serve as the actual target in the experiment. The others are used as controls. The rating values assigned to pictures are based on the agreement between mentation (reported or not) and the content (or perhaps symbolic meaning of the content) of the pictures. Based on the ratings assigned to each response, the pictures can be ranked and one of the familiar evaluation methods for preferential ranking may be applied. But by turning ratings into ranks the greater sensitivity that the rating method might yield is lost. Hence, a statistical evaluation is needed which does credit to the higher sensitivity which ratings might offer. To this end most often Z-scores are applied, first used and reported by Stanford and Mayer in 1974 (JASPR, 1974, 182-191).

When free-response rating data of an experiment were analyzed by applying nonparametric tests on the Stanford Z-score distribution of the targets a significant result indicating psi-missing was observed. However, it soon became clear that the result was purely artifactual and could be explained by the rating behavior of the subjects. This led us to study the properties of the Stanford Z-scores in more detail.

Hansen reported to the 1985 PA Convention (RIP 1985, 93-94) that Z-score distributions are bimodal. We found that Z-score distributions are in all cases nonnormal and only symmetrical but bimodal when subjects select ratings with equal probability from the whole range. Decreasing the size of the target set yields flatter distributions. Decreasing the range of the ratings results in more irregular distributions. All distributions have an upper and lower limit of Z-scores. In cases in which subjects select ratings with unequal probabilities from the range applied, the distributions become asymmetrical. Hence it can be concluded that rating behavior influences the distributions of Stanford Z-scores. This seems an important problem because in many cases the conditions of the experiment will influence the rating behavior of subjects. That implies that an influence of conditions on the rating behavior, and consequently on the Z-scores, must be eliminated before a proper evaluation of the difference as regards ESP scoring can be made.

Stanford Z-scores are also peculiar in some other respects. Their value and range are rather sensitive to the number of equal ratings assigned. In the case in which equal values are assigned, the actual size of the ratings has no influence on the size of the Stanford Z-score. For instance, 1-0-0-0-0 (first rating is target) yields the same Stanford Z-score for the target as 100-0-0-0-0; in both cases the target receives a Stanford Z-score of +1.72. Hence, the Stanford Z-scores do not always reflect the similarities or differences between mentation and targets that subjects express in their assignment of rating values.

Another complication is that when relatively many ratings of equal value are assigned, the Z-score distribution tends to become discrete rather than continuous. Especially since free-response studies in general involve few trials, the discrete character of such distributions violates the assumptions on which many parametric and nonparametric tests are based. To meet these objections a different evaluation procedure based on a randomization test is proposed.

The randomization test is based on the sum of ratings over the trials. In the case of assigning rating values to the control pictures it can be assumed that ESP can have no affect on these ratings. If we randomly select from each trial a control-picture rating value and take the sum of these ratings, then based on all possible combinations of ratings for control pictures over the trials a distribution is obtained which will tend to be normal even in the case that the ratings themselves were selected with unequal probabilities. The randomization test provides an answer to the question to what extent the sum, over the trials, of the ratings assigned to the target pictures deviates from the mean sum of the ratings assigned to the control pictures. Consequently, the sum of the ratings assigned to the target pictures is expressed as a standard normal score based on the distribution of the sum of the ratings assigned to the controls. This standard normal score will be called the "standardized sum-of-ratings score" or SSR score. A good approximation of this distribution is obtained by calculating the mean and standard deviation from the mean and variance of the ratings for the controls of the individual trials. The mean of the distribution of sum of scores will be equal to the sum of the means of controls for the trials. The standard deviation is found by taking the square root of the sum of the variances for control ratings over the trials.

Since SSR scores can be assumed to be standard normal, their associated probability can be obtained from the standard normal distribution. SSR scores of different conditions can be compared because SSR scores not only can be considered standard normal but also are independent of differences between conditions in range of ratings, rating behavior, or number of controls applied.

To obtain ESP scores for individual trials the rating value

assigned to the target is converted into a standardized average rating score for the target (SAR score).

The distribution of the sum of ratings for the controls can be considered as the distribution of ratings associated with that condition. Reduced to the level of individual trials we assume this distribution to be typical for the condition and express all ratings in this distribution of average ratings. Thus, all ratings are converted into standard normal scores by computing its distance from the mean of average ratings for the controls of the trials and dividing it by the standard deviation observed for these average ratings.

Then for each trial a SAR score for the target is defined as the difference between this standard normal score for the target and the average standard normal score for target and controls. Since the SAR scores are based on true standard normal scores, which means scores obtained from a normal distribution, SAR scores can be considered normal too. For each trial the sum of SAR scores for controls and targets is zero. Therefore, in the case of related samples we might compare individual achievement over conditions by calculating a product-moment correlation between the SAR scores of the two conditions.

Although the randomization test described above seems statistically sound we further studied its properties, especially regarding its sensitivity to detect ESP. To this end we conducted a computer simulation of 100 "experiments" for each combination of two variables. Each experiment consisted of 20 trials and 5 pictures per trial and was simulated by randomly generating 20 rows of 5 numbers between rating values 0 and 30, inclusive. The two variables involved were subjects' rating behavior and amount of ESP. For rating behavior we manipulated the probability of selecting rating values of zero. The amount of ESP was operationalized as the number of subjects assigning the highest rating value to the target in addition to what could be expected by chance.

From the data obtained it can be concluded that in most conditions the sensitivity of the SSR scores is rather low and less than that when, for instance, a simple binomial test was applied. Only in extreme cases of rating behavior and amount of ESP do the SSR scores become more sensitive than the binomial test. For instance, in the case of 5 ESP hits when in total 5 + 15/5 = 8 hits can be expected, the binomial yields an exact one-tailed probability of $p_. = .01$ whereas the SSR score yields on average a Z of 1.7 with an associated one-tailed probability of .045.

In the same simulation studies Stanford Z-scores were computed. We know that the distributions for these Z-scores are non-normal but leaving this aside we found that in most cases the sensitivity of t-test evaluations based on Stanford Z-scores is comparable to that of evaluations based on SSR scores. However, SSR

scores appear more sensitive than Stanford Z-scores in cases of strong ESP and extreme rating behavior.

From these findings some practical conclusions can be drawn. In general we must assume that the ESP influence on the data is relatively little. Hence, unless there is reason to expect a strong ESP influence in the experiment the binomial test can be assumed to be more sensitive than an evaluation based on the rating values. The same applies for experiments in which no extreme rating behavior can be expected, for instance, in an experiment in which an atomistic approach to the judging is followed. In that case we expect in general nonzero ratings assigned to all pictures, and our findings show that in that case the SSR scores, as well as Stanford's Z-scores, are rather insensitive.

A METHODOLOGY FOR THE DEVELOPMENT OF A KNOWLEDGE-BASED JUDGING SYSTEM FOR FREE-RESPONSE MATERIALS

Dick J. Bierman (Dept. of Psychology, University of Amsterdam)

It has been found that certain judges perform consistently better than others when matching targets to a target set. It seems unlikely that this is purely because of the judge's psi, since psi generally does not display consistent behavior. Therefore, it might be hypothesized that it is the (intuitive) knowledge of the specific judge that accounts for his better performance on this task. It has been proposed (Morris, ESP, 1986, 137-149) that the use of expert systems might help psi researchers in tasks where they lack expertise, such as in the detection of fraud. Morris argues that the expertise of magicians could be formalized in such a system and made available to each individual researcher. Similarly, the expertise of the best judges of free-response material could become available through implementation of a knowledge-based free-response judging system. This use of techniques from the field of artificial intelligence (AI) to represent scarce knowledge should not be confused with the use of AI techniques for the representation of free-response material (Maren, RIP 1986, 97-99). According to Maren, the free-response material and the protocols should be represented in the form of trees in which the nodes are perceivable "objects," like "flames," and the links represent relations, like "adjacent to." We expect that focusing our attention on the (knowledge used in the) human matching process might reveal more fundamental information about the role of the meaning of the material. It is striking that in Maren's proposed representation of complex target material only visual features are present. Actually, the type of visual matching that Maren proposes to be done by a machine can be better performed by any sighted human.