Originally published in the *Journal of Parapsychology*, 1970, Volume 34, pages 124-128.

THE STACKING EFFECT AND METHODS OF CORRECTING FOR IT

By ROBERT H. THOULESS AND ROBERT M. BRIER[1]

ABSTRACT: Occasionally it is practicable to use an ESP test in which all the subjects make their calls for the same set of targets instead of for individual sets. When a single set is used for the group, it is necessary' to make a statistical correction to compensate for the possibility that the subjects may share the same calling bias and that this may coincide with (or deviate from) the target order and give spuriously high (or low) results (the "stacking effect"). One of the methods of correction was worked out by T. N. E. Greville. Other experimenters have allowed for the stacking effect by merely reducing their critical ratios by 10%, a simplification which lacks theoretical justification. Two other methods pointed out by the authors to evaluate these multiple-calling experiments are the majority-vote technique and the use of the "index of preference."—Ed.

From time to time, experiments are carried out which involve what is called the "stacking effect." In such experiments, it is necessary to compensate for this effect by making a statistical correction in the evaluation of the results. It is the purpose of this paper to describe the stacking effect and to discuss the various methods of correcting for it.

The most efficient design for an ESP test is usually one in which each subject has an individual set of targets. However, there may sometimes be justifiable reasons for adopting the design of an experiment involving a number of subjects guessing at the same target order. For instance, in testing a class of young students, the experimenter may judge that a single target order for all subjects is better motivationally because the experimenter can provide duplicate record sheets and, when the guesses are completed, collect the carbon copies, after which he can read out the targets so that the subject can make an immediate checkup. Another situation in which the single target order is likely to be preferred is when very large numbers of subjects are involved in an experiment.

In a group ESP test in which all subjects have the same target order, the problem of assessing the significance of an observed deviation

[1] This paper was written while the junior author was the Ralph Drake Perry Fellow at the Institute for Parapsychology.

from mean chance expectation in the scores is more difficult than in an ESP test situation in which each subject has a different target order, but the advantages in economizing on experimental time maybe held to outweigh the disadvantages of increased statistical complexity in evaluation.

The ordinary methods used for evaluating results from experiments in which there are individual target orders are relatively easy to carry out, since they depend on the assumption that the result of one run (of a 25-card ESP deck, for instance) is independent of the result of another run. This cannot be assumed, however, when a number of runs of guesses are made for the same target order, for the lack of independence between the results of the runs made by different subjects might produce a stacking effect; that is, a spurious appearance of psi-hitting (or of psi-missing). It might for instance, be the case that an "open" target order (i.e., one in which the frequency of targets in a 25-trial run is randomly determined) might contain more than five of a target which also happened to be a favorite guess. Suppose, for example, that the run contained eight stars and that there was a general tendency for subjects to guess star more often than the other symbols. This would probably result in more hits per run than the mean chance expectation of 5, and this spurious excess over mean chance expectation would, by ordinary methods of calculation, appear more significant as the number of subjects was increased. A spurious deficiency of hits could, of course, equally well be obtained if there were a general tendency among subjects to guess star less frequently than other symbols.

It is not sufficient to avoid this difficulty by substituting a "closed" deck (with five of each kind of target) for the open one. Difficulty would then arise from the possibility that there would be a correspondence between the serial order of the targets and preferred serial orders in the subjects' guessing. For example, suppose that the first two cards in the pack happened to be a square and a circle and that the sequence of square-circle was more commonly chosen for the first two guesses than any of the 24 other pairs that might be chosen. Again, if such preferred sequences, here or elsewhere in the deck, coincided with actual sequences in the target cards, the result would be a spurious tendency for the subjects to score an excess of hits, which might even be magnified to produce an illusory

appearance of significance if enough subjects agreed in their preferred orders of guessing. Also, of course, preferences in guessing sequence may produce an equally illusory appearance of psi-missing.

What is needed to avoid this difficulty is to use a method of evaluating the significance of a set of runs by different subjects on the same target order which is free from the possibility that accidental correspondences between target frequencies and preferred calls or between target sequences and preferred guessing-sequences can produce spurious indications of psi-hitting (or of psi-missing).

The standard way of doing this is by the use of the "Greville correction" of variance for multiple calling developed by T. N. E. Greville and published by him in the *Annals of Mathematical Statistics* (2). Since this publication is not readily available to parapsychologists, Greville's method was presented in simplified form by J. G. Pratt in the *Journal of Parapsychology* (3). It is this form that has generally been used in the past by experimental parapsychologists to deal with the problem of evaluation of results obtained by multiple calling on one deck.

Greville's method involves tabulating the calls made on each target by all the subjects. From this tabulation one can then compute the variance for the calls made on each of the targets. The computation is somewhat different for the open deck and the closed deck; the method for each is given in Pratt's article. For the open deck, when the 25 variances of calls for each card in the target run have been calculated, these are added to obtain the variance for the complete run. The square root of this variance is then taken as the standard deviation to be used in calculating the critical ratio. For the closed deck, the variance for the entire set of 25 targets as the unit must be calculated. Then the square root of this is taken as the standard deviation to be used.

This procedure normally gives a higher variance than the commonly used theoretical variance appropriate for experiments in which only one run of guesses has been made for each target order. Therefore it has the effect of reducing the significance of an observed deviation in either direction. This, of course, is to be expected, since a likely effect of multiple calling is to magnify accidental correspondence or lack of correspondence between target and call.

The tabulation required for the application of the Greville correction

is laborious, and some experimenters have preferred to use simpler methods of arriving at the same result. The simplest was one suggested by Van Busschbach (5). In some of his earlier work, he had applied the Greville correction and had found that the effect was to reduce the critical ratios by less than 10%. This led him to suggest that a 10% reduction of the critical ratios could be used to avoid the complicated calculations required by the Greville method. This simplification lacks theoretical justification; the correction required is not constant for all card distributions or for all numbers of guessing subjects. Osis[2] has discussed cases in which the Greville correction has reduced *CRs* by considerably more than 10%. Estimates of significance based on this 10% reduction, such as those made by Van Busschbach (5) and Brier (1), must be considered to rest on somewhat dubious foundations.

The simplest and most direct alternative to Greville's method would seem to be the repeated-guessing, or the majority-vote, technique in which each subject's guess is treated as a *vote* for the symbol on the corresponding target card (4). The whole system of guesses by all subjects is thus reduced to a single run of 25 guesses. One way of doing this is to take as the final guess on any target card that target symbol which has been guessed by the largest number of experimental subjects. This process gives the experimenter a single final order for the guesses, which he can then compare with the order of the target deck.

A lengthier but more sensitive method of getting this final guess order is by the use of the "index of preference" (4). In this method the votes are weighted in a manner determined by the frequencies of different guesses and of different targets. If a subject (or group of subjects) has a bias for or against calling a particular symbol, this is taken into account when determining which symbol is to be considered the final guess for the group; for example, if one symbol is hardly ever called by the subject and if the subject breaks his habit and makes many calls for this symbol, it might be reasonable to count this symbol as the final guess for the target, even though another symbol might have been called slightly more often.

By this method, the preferred guess for each target card is determined by the use of the formula n_{xy}/gy, where is the number of

[2] In private conversation.

times a certain target card *x* has been guessed as bearing the symbol *y*, and *gy* is the total number of guesses of *y* over all target cards. The preferred guess, then, is that one of the guess symbols which gives the maximum value of this index for each particular target card.

Whichever of the alternatives to the Greville method is used, the result is that, if ESP is affecting a considerable number of the subjects' guesses even to a small extent, the final guess order so obtained will correspond more closely with the target deck than will an individual run of guesses. The evaluation, either by the majority-vote method or the index of preference, is very simple, based as it is on a comparison between a single target order and a single set of guesses. It gives an estimate of the significance of the correspondence between target order and subjects' guesses which is entirely free from any effect of magnification by multiple guessing due either to commonly preferred targets or commonly preferred sequences. It achieves the end of the Greville correction (for either an open or closed deck situation) in a simpler and somewhat less laborious way.

One obvious suggestion that may follow from the above discussion is that experiments should be designed, when at all possible, so that an individual set of targets is presented for each subject. This would eliminate the need for considerations such as those discussed in this paper.

REFERENCES

1. BRIER, R. M. A correspondence ESP experiment with high-I.Q. subjects. *J. Parapsychol.,* 1967, **31**, 143-48.
2. GREVILLE, T. N. E. On multiple matching with one variance deck. *Annals of Mathematical Statistics,* 1944, **15**, 432-34.
3. PRATT, J. G. The variance for multiple-calling ESP data. *J. Parapsychol.,* 1954, 18, 37-40.
4. THOULESS, R. H. The repeated-guessing technique. *Int. J. Parapsychol,,* 1960, **2** (No. 3), 21-36.
5. VAN BUSSCHBACH, J. G. A further report on an investigation of ESP in school children. *J. Parapsychol.,* 1955, **19**, 73-81.

2 Leys Rd.
Cambridge CB4 2AU
England

Institute for Parapsychology
College Station
Durham, N. C. 27708