THE JOURNAL OF THE AMERICAN SOCIETY FOR PSYCHICAL RESEARCH

VOLUME 76 APRIL 1982 NUMBER 2

On Grouping of Hits in Some Exceptional Psi Performers

EDWARD F. KELLY[1]

ABSTRACT: The data of 10 series of forced-choice experiments involving a variety of psi tasks and six high-scoring subjects were examined for nonrandomness in the within-run grouping of their hits. For two subjects, grouping analyses were also made of parallel visual series in which they tried to identify briefly-displayed slides of the same targets used in the psi tasks. Five of the six subjects showed significant evidence of nonrandom grouping in their psi data. In four cases the effect took the form of an excess of long clusters of hits, while the remaining case showed significant isolation of hits. Grouping effects were also observed in both visual series, of a form and magnitude comparable to those of the parallel psi series.

The grouping effects in the psi data appear not to arise from external parameters such as type of task, trial rate, feedback regime, etc., but rather to reflect individually patterned tendencies toward transient establishment of conditions favorable for successful psi response. These effects accounted for much of the evidence for psi in the experiments in which they occurred, and in some cases appear to have been reflected in the subjects' physiological measures, their subjective states, or both. Implications of these findings for further work on psi-favorable states are sketched.

INTRODUCTION

Historically speaking, evidence of psi in the results of parapsychological experiments has most often been assessed primarily in terms of the overall rate of occurrence of hits, direct or displaced. However, many additional structural features of data from typical psi experiments are accessible to quantitative study (as reviewed, for example, by Burdick and Kelly, 1977) and, particularly in extended series of tests with individual subjects, some of these

[1] I wish to express my appreciation to the McDonnell Foundation, St. Louis, Missouri, for financial support; and to Donald S. Burdick, Irvin L. Child, and Ralph G. Locke for constructive criticisms of an earlier version of this paper.

less-frequently studied features hold far more promise than does the raw hit rate of providing significant insight into psychological aspects of the performance. One prime example here is "consistent missing," or systematic erroneous association of particular targets and responses; another is sequential grouping of hits. It is no coincidence that the study of both these topics was pioneered by the late J. G. Pratt (Cadoret and Pratt, 1950; Pratt, 1947).

In his 1947 paper, Pratt introduced a method developed by Stevens (1939) which permits statistical evaluation, independent of the scoring rate, of the tendency for hits to occur in groups. Given the number of hits occurring in a sequence of trials, the Stevens method allows us to determine whether those hits are distributed nonrandomly within the sequence. Thus, to a considerable degree the question of *occurrence* of psi is analytically separated from an important question about its *characteristics*—namely, whether there is evidence of trial-to-trial continuity in whatever conditions have led to correct psi responses.

To address this question, Pratt applied tests for grouping to seven already-existing bodies of psi data selected to represent a wide range of experimental conditions and scoring levels. These included two dice-throwing PK series (Gibson cup and Reeves high-dice and low-dice) and five ESP series (Riess, Pearce-Pratt distance, Pearce clairvoyance without distance, Soal-Goldney [82 runs with Shackleton], and the Ownbey-Turner telepathy series). For comparison with the ESP results, Pratt also analyzed visual data collected by Burke Smith from eight subjects who briefly viewed ESP cards in dim light. He did not find any overall evidence of nonrandom distribution of hits in the psi series. However, he noted that in two of these, the Pearce-Pratt distance series and the Reeves high-low dice, the high-scoring portions showed moderately significant tendencies toward nonrandom *isolation*, rather than clustering, of hits. The visual data, by contrast, showed strong evidence of clustering, and that pattern held throughout the range of run scores. Pratt tentatively interpreted these exploratory results as consistent with a model in which the psi process operates instantaneously and at an unconscious level in conjunction with single trials.

To my knowledge, the subsequent experimental literature contains only three reports of analyses for stringing or grouping of psi hits. Soal and Bateman (1954, Appendix D) reported an analysis of the complete sequence of 37,100 GESP trials with Gloria Stewart, giving both the overall Stevens test for grouping and an approximate analysis of observed and expected numbers of hit-strings of length 1, 2, . . ., 12. The results showed an extremely significant grouping effect ($Z = 7.94$, $p < 10^{-14}$), which the authors interpreted

as a secondary consequence of Mrs. Stewart's strong tendency to crowd her hits into the first three-fifths of the run (p. 314). Musso and Granero (1973, 1981) reported strong clustering of hits in a highly successful single-subject free-response experiment, and showed that this was due to the tendency for the results of individual six-trial sessions to be consistently either strong or weak.[2] Schouten and Kelly (1978) reported briefly on clustering of hits in the Brugmans experiment with Van Dam, noting a very strong clustering effect that appeared to be due partly, but not entirely, to a few outstanding sessions in which Van Dam produced long sequences of direct hits on the target squares.

These latter results clearly tend in quite a different direction from those of Pratt (1947), and invite further systematic study of grouping effects in an increased variety of subjects and psi tasks. The purpose of this paper is to report the results of one such investigation.

METHOD

Data

In all, 12 sets of data (10 psi series and two visual series) representing six exceptional subjects were used for this study. Some of these datasets have been analyzed for grouping effects at least in part in earlier studies, but most have not, and in all cases new information is provided. The first three datasets were generously made available by J. G. Pratt in the form of computer-ready decks of keypunched cards. The subjects and datasets analyzed are as follows:

1. *Miss S* (the subject of the Riess (1937, 1939) experiment). The data consist of 74 25-trial runs of GESP with the standard ESP deck, called at a rate of one card per minute at a distance of a quarter mile with no feedback at any time (Riess, 1937, p. 262). Miss S averaged over 18 hits per run for the entire series.

2. *Hubert Pearce*. The data are those of the Pearce-Pratt distance series (Rhine and Pratt, 1954), also previously analyzed for grouping by Pratt (1947), consisting of 74 25-trial runs of clairvoyance on ESP card decks with calls spaced at one per minute and feedback at the end of sessions (one to three runs). Four subseries were carried out, with the first, third, and fourth at a distance of 100 yards and the second at 250 yards. Pearce averaged over 7.5 hits per run for the entire experiment.

3. *Gloria Stewart*. The data analyzed come from Series I, II, and III of the Soal-Stewart data as described in Pratt (1967, pp. 30–32).

---

[2] My thanks to Irvin Child for bringing this result to my attention.

Briefly, these records include the bulk of Mrs. Stewart's high-scoring work between August 1945 and late 1947 under Soal's standard GESP testing regime. Data from runs under "abnormal" conditions—for example, clairvoyance, binary targets, unusual calling-rates, senders in opposition, etc.—were systematically excluded according to principles stated by Soal and Pratt (1951). Series I and II, each consisting of 212 runs of 25 calls, utilized the G, E, L, P, Z targets (giraffe, elephant, lion, pelican, and zebra), but Series II terminated at a point just before Mrs. Stewart's scoring fell to chance with these targets. The experimenters attributed this decline to her having begun mechanically spelling out three-letter words and nonsense syllables, and consequently they changed the targets early in 1947 to C, F, H, K, T (camel, fox, horse, kangaroo, and tiger), to exclude the initial vowel. Her high scoring promptly resumed, and continued until mid-1949. Series III consisted of 200 runs from the beginning of this period. I do not have copies of the additional series SS-IV, SS-V, SS-VI, and PS described by Pratt (1967). The data analyzed nevertheless collectively represent a considerable sample of Mrs. Stewart's high-scoring performance, collected under fairly homogeneous experimental conditions. Although the variability of run-scores has been reduced at the low end by selection, the sample still contains a substantial number of below-chance runs, and in any event this selection has no bearing on the main analysis for stringing effects, which takes run scores into account as indicated above.[3]

4. *Van Dam*, the single subject of the experiment of Brugmans, Heymans, and Weinberg (Schouten and Kelly, 1978). The complete experiment consisted of 589 trials organized in 24 sessions of varying length. The subject's task was to identify the target square in a 6 x 8 checker-board-style array, which he did successfully on 118 occasions (for grouping analyses I have used the dataset based upon uniformly conservative resolutions of ambiguous responses, as described in Schouten and Kelly [1978, p. 203]).

5. *Lalsingh (Sean) Harribance*. Two series involving Zener

[3] Some readers may wonder whether it is appropriate to include data collected by S. G. Soal, given the cloud that has fallen over his reputation due to the series of recent critical attacks by Markwick (1978) and others. I offer three reasons for inclusion: First, particularly in the light of a recent conversation with Evan Harris Walker, I am not convinced that the controversy over the Soal-Shackleton data has been resolved conclusively and against Soal. Second, such explicit evidence as there is against Soal is so far confined to the Shackleton series. Third, cheating of the form proposed for the Shackleton series would not, I believe, lead to grouping effects, and indeed none were reported for it by Pratt (1947). Hence, if the strong clustering effects in the Stewart data are to be explained by cheating, it will have to be of a quite different sort. Nonetheless, I agree that the results reported here should be regarded cautiously in the light of the present status of this controversy.

cards have been analyzed. The first is a sample of 50 ESP runs, each consisting of 25 down-through clairvoyance calls, randomly selected from a 445-run subset (for which listings were available) of the high-scoring 1000-run series reported by Morris (1972). The sample contained 310 hits in 1250 trials ($Z = 4.2$). The conditions of the experiment, while deliberately informal, were good enough to create a strong presumption that the results are representative of Sean's ESP performance. Since, like Pratt, I am interested in comparisons between normal and paranormal processes, the second series chosen for Sean is a visual series in which he attempted to identify briefly-projected slides of ESP cards (Kelly, Child, and Kanthamani, 1974). The series consisted of 25 runs of 50 calls and contained 671 hits, the high scoring rate (54%) creating a reasonable presumption that any internal effects in the visual series were contributed primarily by visual rather than psi factors.

6. *Bill Delmore.* Six bodies of data available from previous experimental work with B.D. and representing a variety of tasks have been studied for grouping effects. The first group of four datasets includes all of B.D.'s work involving playing cards (except for a few very short series involving special tasks, the results of which have not been published). Set one consists of 73 runs of 52 calls of trial-by-trial clairvoyance with immediate feedback, conducted by Irvin Child, in which only marginal evidence of psi was obtained. Set two consists of 46 runs of 52 calls, using the same basic procedures, conducted primarily by H. Kanthamani. Although feedback was usually available on each trial, groups of non-feedback trials were occasionally interspersed at B.D.'s request. This series contained extremely strong evidence of psi-hitting at the level of number-hits and higher, as well as evidence for a pattern of systematic errors analogous to those he made in a corresponding visual task. Sets one and two are described in Kelly, Kanthamani, Child, and Young (1975) and Kanthamani and Kelly (1974a). Set three consists of 55 runs in which B.D. shuffled a deck of playing cards to match a prearranged target deck. Extremely strong evidence of psi was again obtained, but the effect was concentrated in an excess of exact hits and there was no indication of the pattern of near-misses displayed by set two. Another important difference is that the psi effects in the shuffles series appeared to be primarily psychokinetic in type because of the small numbers of shuffles B.D. needed to produce large numbers of hits (Kanthamani and Kelly, 1975). Set four is a visual series consisting of 75 runs of 52 calls in which B.D. attempted to identify briefly projected slides of playing cards. As in the Harribance visual series, the scoring rate was high enough (60% exact hits) to make it highly likely that any internal effects it may contain were due primarily to visual rather

than psi factors. Note that in the present study, in contrast to that of Pratt (1947), comparisons between visual and psi effects can be based upon psi and visual data collected from the *same* individuals. In B.D.'s case, these data were also collected during the same period of time.

A second group of two series involves B.D.'s work with Schmidt 4-choice machines. Set one is the long series reported in Kelly and Kanthamani (1972). In using our laboratory computer to read the punched paper output tapes from this series for purposes of the present study, I found a few minor discrepancies from the manual counts previously published, and one block of 43 trials of unrecorded origin. Although this latter set of trials may or may not have been produced by B.D., I have included it in the interests of conservatism. The result is a total series of 5425 trials in eight sessions of varying length, in which B.D. scored 1549 hits or 28.5% ($Z = 6.06$). During this entire series the machine was operated in the PK mode; i.e., if and only if a "1" target was generated, the circuitry would count a hit and light the feedback lamp corresponding to whichever response button had been chosen. Set two is a previously unreported series carried out by Irvin Child in New Haven during May 1972, using a Schmidt machine operating in the "normal" (precognition?) mode, and consisting of 1800 trials organized in 100-trial runs scattered over three testing sessions. Although this series was not taped, its records included the trial-by-trial sequence of hits and misses, and thus permitted analysis for grouping. In these 1800 trials, B.D. scored 501 hits or 27.8% ($Z = 2.74$, $p<.01$).

Analysis Methods

The basic statistical methods available for analysis of grouping effects are outlined in Burdick and Kelly (1977, pp. 105–108). For the statistical test of grouping I have used the method of Wald and Wolfowitz (1940)—henceforth, W-W—in preference to that of Stevens (1939). Stevens' method is based on groups of *successes* only, whereas the W-W method is based upon groups of both kinds; but the associated distributions are essentially the same, and in particular both are determined by the *observed* number of hits in the sequence to be analyzed. The W-W method is more widely known, however, and critical values of the total number of groups, d, have been tabulated for cases in which the number of successes, m, and the number of failures, n, are both less than 20 (Siegel, 1956; Swed and Eisenhart, 1943).

In order to generate more information about the *form* of any grouping effects, additional computer routines were developed to

calculate the expected numbers of hit-groups of length 1, 2, . . ., and to count the corresponding numbers of observed groups. It should be emphasized that the calculation of these expected values, like the statistical test for grouping itself, is based upon the *observed* number of hits in the sequence, so that grouping effects are analytically distinguished from secondary effects of altered hit rate (Burdick and Kelly, 1977, pp. 107–108).

The main analyses for grouping effects were carried out by applying these procedures to individual runs of each series and accumulating totals across runs. For convenience, the computational details are provided in Appendix A.

Particularly where series consisted of runs of uniform length, secondary analyses were carried out to explore the relationship, across runs, between overall scoring level and grouping effects. These analyses took several forms. First, the W-W results could be accumulated separately for each run-score, and then for subgroups of runs as appropriate, providing a clean statistical test of grouping effects in each subgroup. A test of the difference in grouping effects between two batches of runs could then be constructed from their individual W-W results Z_1 and Z_2 by using $Z_{\text{diff}} = (Z_1\text{-}Z_2)/\sqrt{2}$. Of particular interest here is the contrast between low-scoring and high-scoring runs.

The remaining procedures took the individual run as the unit of analysis, and characterized the amount of grouping within it by its Z-score calculated from the W-W test. Then relationships between overall run-scores and grouping effects could be explored using standard devices such as scatterplots, Spearman rank correlations, and analysis of variance for differences in grouping effects among batches of runs bracketed in terms of overall score.

These run-level analyses, however, are slightly impaired by a technical difficulty which should be mentioned here. In effect, in doing such analyses we are weighting all runs equally, but the *Z-score for grouping is not equally dependable as an index of grouping at all scoring levels. The* W-W "Z-score" of course only provides a normal approximation to an underlying discrete probability distribution. Wald and Wolfowitz (1940) showed that the distribution of the number of runs is asymptotically normal, and Mood and Graybill (1963) remark that in fact the normal approximation "is usually good enough for practical purposes when both m and n exceed 10" (p. 411). (Note also that for Mood and Graybill "practical purposes" meant *estimation of probabilities,* whereas we are only requiring of these run-level Z-scores that they provide a reasonably accurate index of the amount of grouping in the run.) When sequences are short and/or p(hit) is extreme, however, cases are generated lying outside this range. Most conspicuously, in

B.D.'s playing-card psi data a substantial number of runs occur containing two or fewer exact hits in 52 trials. Runs with 0 or one hit are essentially irrelevant to the question of grouping, so these can reasonably be excluded. Runs with two hits, however, exemplify the problem very well: The only possible outcomes in this case are 2, 3, 4, or 5 groups of hits or misses, and five is overwhelmingly the most likely. The exact mean and variance for d are in fact 4.85 and .21, respectively. Thus if five groups occur, mechanical computation of the Z-score leads to a result, $-.75$, which is dominated by the continuity correction and artificial in both sign and magnitude. Fortunately, this problem rapidly dissipates as the number of hits increases, leaving only a small residue of fuzziness in the neighborhood of $Z = 0$. For the few more extreme cases such as the above, I adopted the simple but arbitrary policy of setting the Z-score to zero. This has the effect of bringing the Z-scores that result from such extreme-value cases into reasonable conformity with what we intuitively want the scores to mean for purposes of further analysis. In a few cases I also tried alternative ways of handling the problem, which seemed to make very little difference and were considerably more laborious. Thus, although the inaccuracy of the normal approximation in extreme-value cases certainly introduces a small amount of error to the data, particularly in the vicinity of $Z = 0$, this error does not appear to be systematic or to have any major bearing upon the results. I therefore conclude that use of the run-by-run Z-scores in the ways described is adequately justified for the exploratory purposes of this report.

Additional analytical devices specialized to particular bodies of data will be introduced subsequently at the appropriate points.

Results

I will now present the main results of the grouping analyses for each subject in turn, deferring interpretations for the most part to the discussion section.

Miss S. Results of the overall analysis of grouping in the 74 runs of the Riess series are presented in Table 1, Part A. The W-W test shows that the observed total number, d, of groups of hits or misses is right at MCE, so there is no evidence whatsoever of grouping. Although our numbers differ slightly, this is essentially the result previously reported by Pratt (1947).

The expected and observed distributions of Miss S's hits into strings of progressively greater length are presented in Part B, truncated at the length of the longest observed string of hits (omitting the perfect run 19). Note that the "expected" and "observed"

Table 1

OVERALL GROUPING RESULTS FOR MISS S (RIESS SERIES)

A. Wald-Wolfowitz Test

# Runs	d	E[d]	Dev	% Dev	σ^2 (d)	Z
74	714	714.5	0	0	229.9	0

B. Hit-Group Data[a]

Length of Group	Expected #	Observed #
1	117.27	114
2	74.60	67
3	50.31	64
4	35.14	34
5	25.09	25
6	18.17	20
7	13.29	10
8	9.79	5
9	7.25	4
10	5.39	4
11	4.01	6
12	3.00	6
13	2.24	6
14	1.68	0
15	1.27	2
16	.97	3
17	.76	0
18	.60	0
19	.50	0
20	.42	0
21	.38	0
22	.35	1

[a] Run 19, with a perfect score of 25, was eliminated from this analysis.

columns simply partition the same total number of hits in different ways: that is, for each column the sum, over successive lengths, of [(length) × (number)] equals the total hit-count. Thus the rows are not independent, since deficits must be compensated by excesses elsewhere. Although it is not valid to apply the standard chi-square goodness-of-fit test to these tables, they are very helpful in quickly characterizing the form of grouping effects, as will become more apparent in the sequel.

Although Miss S's data show no overall tendency toward grouping, the observed distribution of string-lengths appears suspiciously choppy, with quite large fluctuations of observed relative to expected numbers at successive lengths. For example, a modest deficit of lengths 1 and 2 is more than compensated by a considerable excess at length 3, and 7–9 are substantially under-represented while 11–13 are at least equally over-represented. In each of these

Table 2

GROUPING RESULTS FOR HUBERT PEARCE (PEARCE-PRATT SERIES)

A. Wald-Wolfowitz Test

# Runs	d	E[d]	Dev	% Dev	σ^2 (d)	Z	p[a]
74	829	778	+51	6.55%	283.3	3.03	.003

B. Hit-Group Data

Length of Group	Expected #	Observed #
1	253.06	296
2	79.52	73
3	27.13	21
4	9.52	3
5	3.29	3
6	1.09	2
7	.34	2
8	.10	0
9	.02	0
10	.01	0

[a] All *p*-values for these tests are two-tailed unless otherwise stated.

two regions the observed distribution is shifted toward greater lengths, suggesting the possibility of a mixture of weak grouping effects corresponding to different scoring levels. To investigate this, I recomputed the W-W analysis separately at each run-score, but again failed to find clear evidence of grouping. Hence, despite its suspicious appearance, this odd pattern seems on present indications best regarded as the product of random variation.

Hubert Pearce. Results of the overall analysis for the 74 runs of the Pearce-Pratt distance series are presented in Table 2. The W-W test shows a significant tendency toward *isolation* of hits, since there is a large positive deviation in the observed number of strings. This is clearly reflected in the hit-group data, which are principally marked by a large excess of singletons and depletion of groups of lengths 2, 3, and 4.

Assuming that the present results are correct,[4] the next question

[4] These results are substantially stronger than those originally reported for this series by Pratt (1947, p. 260) using the Stevens method. In an effort to resolve this discrepancy, I re-analyzed the data manually according to the Stevens method. The first step was to check the table of means and variances for the number of success groups given by Pratt (1947, p. 267), which proved to be completely correct. Using these figures in conjunction with the run-score distribution, I next calculated the overall mean and variance for the series. The mean agreed with Pratt's results, which not only verifies this figure itself, but indicates that the run-score distributions for raw and keypunched data are identical. Pratt's standard deviation, however, is slightly inflated (8.80 vs. 8.55). The main discrepancy proved to lie in the

concerns the relationship of the isolation effect to overall scoring level. Following Pratt's lead, I divided the data into two groups—36 runs with scores of 9 or higher (10.72 hits/run) and 38 runs with scores of 8 or less (4.53 hits/run)—and recomputed the W-W tests. For the low runs the result was $Z = .18$, while for the high runs it was $Z = 3.40$ ($p < .001$). The difference between these two results is also significant ($Z_{diff} = 2.28$, $p < .02$). The effect in the high-scoring runs is very sizeable as well as statistically significant, with a percent deviation (100 × deviation/expected number) of +10.4 in the observed value of d (the corresponding value for low-scoring runs is only .5%). Thus, the isolation effect in Pearce's data is very largely concentrated in his high-scoring runs. This result, like the overall result, is essentially an amplified version of what Pratt (1947) had already reported.

Gloria Stewart. Soal and Bateman (1954, Appendix D) presented data showing an extremely significant grouping effect when Mrs. Stewart's 37,100 "normal" GESP trials were considered as one continuous sequence. Soal and Bateman apparently believed that this result was simply a necessary consequence of the fact that the aggregate data showed considerable inhomogeneity of hit rate across fifths of the run, with a marked drop on trials 16–25. Thus they remarked that "this crowding of hits into the first three segments of the run is undoubtedly responsible for the wide divergence of the numbers of runs of 1, 2, 3, 4, or more consecutive hits from their expected values" (p. 314).

This conclusion, however, is rather too hasty. That their inference, although at first glance plausible, is not necessarily correct had already been demonstrated by Pratt (1947), who showed in the case of the Gibson cup dice-throwing series that strong position effects in the *aggregate* data were not reflected at the level of the individual runs, and thus could coexist with perfect absence of within-run grouping effects. It should also be pointed out that on Soal and Bateman's own principles, their decision to analyze the series as one long sequence risked confounding grouping effects due to *run-position* inhomogeneity with grouping effects due to *series-position* inhomogeneity, since the latter is certainly also present in Mrs. Stewart's data.

Accordingly, I have reanalyzed the data from the three available

---

observed number of success groups, which Pratt apparently undercounted by 14 (this is easily understandable if he was working manually from raw records). When these adjustments are made, the Stevens and the W-W results are identical apart from rounding error, as they should be. It remains possible, though unlikely, that the keypunched data differ from the raw data in ways that account for these discrepancies; however, one would have to compare them directly to be sure, and so far I have been unable to locate a copy of the original records.

series with two modifications of Soal and Bateman's procedure. The initial analysis was carried out within runs, which effectively eliminates any effect of position within the series, and makes the unit of analysis equivalent to the unit of test performance. I then repeated the grouping analysis within the first 15 trials of each run; this eliminates most of the effects of within-run inhomogeneity as expressed in the aggregate data, since the overall scoring rates in the first three segments of the run are not distinct by a chi-square test.

The main results of these analyses are presented in Table 3. We first note that, allowing for the reduced amount of data, the results of the full-run analysis are substantially identical in pattern to those originally presented by Soal and Bateman. Thus, inhomogeneity of scoring within (and across) series evidently had little effect on the results.

Much more surprising and interesting, however, is the fact that the 15-trial analysis also yields very comparable results. This shows clearly that there is a genuine within-run grouping-of-hits effect in Mrs. Stewart's data which is *not* the secondary result of the run-position effect observed in the aggregate data. Inspection of Part B of Table 3 shows that this effect is characterized by consistently large excesses of groups of three or more hits. (Note, however, that the overall effect, expressed as percent deviation, is actually somewhat smaller in magnitude than the effect we saw in the Pearce-Pratt data, although statistically more significant because of the considerably larger number of runs.)

The next step was again to examine the relationship between the grouping effect and the overall level of scoring in the run. Because of the amount of labor involved, I used only Series II, where the grouping effect appeared most strongly. The 212 runs were divided into two sets, 92 runs with scores of 1 to 6 (4.78 hits/run), and 120 runs with scores of 7 to 13 (8.54 hits/run), and the W-W analyses computed separately for each run-score and set. For the low runs the result is $Z = -3.28$, $p < .002$, while for the high runs it is $Z = -5.93$, $p < 10^{-9}$. Thus there is strong evidence in this series for grouping in *both* high-scoring and low-scoring runs, although the effect is again considerably larger in the high-scoring runs. Ten run-scores (3 through 12) occurred often enough to justify individual tests for grouping, and of the resulting W-W tests, seven are independently significant (two of the four for low scores and five of the six for high scores), and all are in the same direction. For the 35 runs with scores of 10 or higher, the weighted average deficit in the number of strings, d, amounts to 15.6%. An association between run-score and grouping is further indicated by a Spearman rank

Table 3

GROUPING RESULTS FOR GLORIA STEWART (SOAL-STEWART DATA, SERIES I-III)

A. Wald-Wolfowitz Tests

	Full 25-Trial Runs						First 15 Trials Only					
Series	d	E[d]	% Dev	σ^2 (d)	Z	p	d	E[d]	% Dev	σ^2 (d)	Z	p
I (212 runs)	2167	2244	−3.43%	761.2	−2.78	.006	1424	1463	−2.67%	471.3	−1.78	.075
II (212 runs)	2052	2240	−8.39%	763.9	−6.79	10^{-9}	1307	1419	−7.89%	443.7	−5.29	10^{-7}
III (200 runs)	1944	2020	−3.76%	655.0	−2.96	.004	1248	1313	−4.95%	402.8	−3.22	.002
Totals (624)	6163	6504	−5.24%	2180.1	−7.29	10^{-12}	3979	4195	−5.15%	1317.9	−5.94	10^{-9}

B. Hit-Group Data (Aggregated Across Series)

Length of Group	Expected #	Observed #	Expected #	Observed #
1	2300.91	2092	1426.03	1279
2	597.42	535	391.33	344
3	155.46	195	107.80	141
4	40.44	55	29.71	37
5	10.48	21	8.24	18
6	2.69	4	2.35	3
7	.67	11	.71	6
8	.16	7	.23	1
9	.03	2	.07	2
10	.00	0	.01	0

correlation of −.139 (higher run scores paired with more negative Z-scores for grouping), $t = -2.03$, $p < .05$.

The apparently strong evidence of grouping in the "low-scoring" runs is somewhat puzzling, but may be at least partially an artifact of data selection, since this series as a whole was chosen to represent Mrs. Stewart's high-scoring performance, and terminates (as noted above) before she fell to chance with the GELPZ targets. Thus it may be that what are here described as low-scoring runs are really runs from the extreme low end of the chance distribution into which modest numbers of ESP hits have been injected, rather than "true" chance runs. To explore this further, it would be of great interest to examine segments of Mrs. Stewart's data in which the *overall* results were at chance, but unfortunately these records were not available.

It could also be of real interest to examine fluctuations of the grouping effect in relation to additional breakdowns of the presently available data—for example, A vs. B columns of pages, different pages within a session, chronological position of the run, etc.; however, these analyses did not appear sufficiently relevant to the central purposes of this report and have not been carried out.

Before passing to the next subject, let me again underline the sharp contrast between these strong grouping results for Mrs. Stewart and the complete absence of grouping results reported by Pratt for 82 runs of the Soal/Shackleton series. Furthermore, if Pratt made any undercounting errors analogous to those he apparently made in analyzing Pearce's data, the true contrast between the Shackleton and Stewart results would be that much stronger.

Van Dam. The experiment of Brugmans, Heymans, and Weinberg presents a somewhat different analysis problem, both because of the unequal and indeed widely varying lengths of its 24 runs/sessions (six to 47 trials), and because of the multiple-aspect nature of its targets, which can be analyzed for hits on the letter (column) attribute, the number (row) attribute, or whole squares. The attribute tests are of course not independent of the main tests on the squares.

Schouten and Kelly (1978) reported a strong grouping effect in Van Dam's results, based upon an analysis which treated the 587 usable trials of the experiment as a single continuous sequence. We noted that this effect was apparently due in substantial part to a few extremely strong early sessions containing long strings of hits, but that clustering of hits seemed to be present in the remaining trials as well, although the groups became progressively smaller and more scattered.

To increase the precision and detail of the Van Dam results I have now repeated the original analyses, adding the string-length

data, and also reanalyzed the entire experiment on a session-by-session basis to remove most of the effect of series inhomogeneity in the scoring rate.

The main results are presented in Table 4. Clearly, the session-by-session analysis sharply reduces the apparent magnitude and statistical significance of the grouping effect, although it is still present, particularly in the number-attribute data. In large part this is due to the fact that the likelihood of the three very long groups of hits (strings of 12, 8, and 6 exact hits in sessions 3, 2, and 12, respectively) is now being measured in the context of the runs in which they occurred, rather than in the context of the entire series. Thus, for example, the group of 12 hits in run three is no longer so remarkable, since the run contained 16 hits in 18 trials; in fact, this one run contributes most of the shift of expectations toward longer string-lengths in the right-hand side of Table 4, Part B.

Examination of the hit-group data nonetheless reveals that runs 2, 3, and 12 have still contributed strongly to the remaining session-by-session grouping effect for hits on squares.[5] This is particularly true of sessions 2 and 12, where the groups of hits occurred in the midst of substantial numbers of misses. Deletion of these runs in fact causes the session-by-session W-W results to fall to chance for hits on squares, although they remain highly significant for hits on numbers ($Z = -2.93$, $p < .004$, -11.75% deviation).

Although the grouping effect for hits on squares is thus somewhat more dependent on a few key runs than Schouten and I had realized at the time of our 1978 report, it remains significant under the more detailed analyses. The grouping effect is also even more strongly present for number-hits.

Finally, it should be underscored that these effects also appear to be systematically underestimated due to the form of the tests applied, which dichotomize the data and give no credit for near-misses. Inspection of the raw data reveals that exact hits tend to be flanked on one or both sides by responses that miss their targets by just one or two squares. An interesting case in point is run 4, which contains nine apparently isolated hits in a sequence of 36 trials, corresponding to a Z-score for stringing of $+1.37$. This is not statistically significant, of course, but it constitutes the single most conspicuous exception to the general trend of the data. Yet, inspection of the raw records for this run shows that it contains no

---

[5] Readers may note two discrepancies between the *Observed* columns on the left and right sides of Table 4, Part B. One hit-group of length 2 in the complete data fell on the boundary between sessions 4 and 5, and thus was counted as two groups of length 1 for the session-by-session analysis; and one hit occurred by itself in run 24, which caused that run to be omitted from the session-by-session hit-group analysis.

Table 4

GROUPING RESULTS FOR VAN DAM (BRUGMANS, HEYMANS, AND WEINBERG)

A. Wald-Wolfowitz Tests

	Whole Series						Session-by-Session					
	d	E[d]	% Dev	σ^2 (d)	Z	p	d	E[d]	% Dev	σ^2 (d)	Z	p
Letters	238	270.2	−11.9%	123.2	−2.85	.005	256	259.7	−1.4%	94.1	−.33	n.s.
Numbers	187	245.8	−23.9%	101.9	−5.78	10^{-8}	201	239.9	−16.2%	79.3	−4.31	.00002
Squares	149	189.6	−21.4%	60.4	−5.16	10^{-7}	165	182.1	−9.4%	47.3	−2.41	.02

B. Hit-Group Data (Squares Only)

	Whole Series		Session-by-Session	
Length of Group	Expected #	Observed #	Expected #	Observed #
1	75.6166	54	64.1542	55
2	15.1233	13	12.7430	12
3	3.0039	4	3.0217	4
4	.5925	0	.9376	0
5	.1161	0	.4155	0
6	.0226	1	.2627	1
7	.0044	0	.2076	0
8	.0008	1	.1791	1
9	.0002	0	.1574	0
10	.0000	0	.1374	0
11	.0000	0	.1176	0
12	.0000	1	.0980	1
13			.0784	
14			.0588	
15			.0392	
16			.0196	

less than 10 additional responses that fall within a single square of the target, all 10 being consecutively connected to at least one of the nine exact hits. For example, five are contiguous with hits number 3 and 4, making a chain of seven consecutive excellent responses. We could of course seek to take account of such events by liberalizing our criterion for a hit to include misses of arbitrary proximity and recomputing the grouping analyses. We could also consider representing each response by its distance (= $\sqrt{[}$ letter distance^2 + number distance^2]) from the target, and utilizing some sort of serial correlation precedure, successive squared differences, etc. For present purposes, however, we have carried this analysis far enough.

As mentioned previously, the sessions of this experiment were extremely variable in length. Moreover, the decision to terminate or continue a session after a given trial was apparently often based upon the results of the session up to that point. Therefore I have not attempted a formal statistical analysis of the relation between scoring and grouping. Nevertheless, informally at least it appears likely that such a relation did hold, with grouping of hits and unusually strong scoring again tending to occur together.

Lalsingh (Sean) Harribance. Overall results of the main grouping analyses for both the ESP series and the visual series with Sean are presented in Table 5. The W-W tests lead to quite similar results for the two series, with strong evidence for grouping of hits in both. The *pattern* of the hit-group data is also similar for the two series, bearing in mind that for the visual series the runs were twice as long (50 trials vs. 25) and the hit rate very much higher (53.7% vs. 24.8%). Both series show deficits, relative to expectation, of hit-groups of short lengths, compensated by a fairly uniform excess of groups of greater length. For the ESP series, a clear crossover apparently occurs between lengths 3 and 4, whereas in the visual series crossover seems, less clearly, to occur between lengths 6 and 7.

There is a striking difference, however, in the way in which these grouping effects are distributed through the two series in relation to scoring levels in the runs. In the visual series, the tendency toward clustering of hits is present in both high-scoring and low-scoring runs, but with a trend toward more clustering in lower-scoring runs. For example, dividing the data into 12 "low" runs (scores of 14–25) and 13 highs (26–36) yields W-W Z-scores of -2.92 for the low runs and -1.54 for the high runs. The Spearman rank correlation for 25 runs is $r_s = +.158$, with $t_{23} = .77$, indicating again a nonsignificant trend toward less grouping at higher run scores. Similarly, ANOVA of high vs. low runs yields ${}_1F_{23} = .75$, which is also nonsignificant. These results are consistent with those re-

Table 5

GROUPING RESULTS FOR SEAN HARRIBANCE

A. Wald-Wolfowitz Test

Series	# Runs	Trials/Run	d	E[d]	Dev	% Dev	σ^2(d)	Z	p
ESP	50	25	442	480.7	−38.7	−8.05%	148.9	−3.13	.002
Visual	25	50	556	608.8	−52.8	−8.67%	268.3	−3.19	.002

B. Hit-Group Data

	ESP Series		Visual Series	
Length of Group	Expected #	Observed #	Expected #	Observed #
1	170.82	165	145.54	133
2	40.65	29	73.24	66
3	10.86	7	38.11	28
4	3.37	5	20.50	17
5	1.21	4	11.39	12
6	.49	3	6.53	3
7	.21	0	3.85	4
8	.09	1	2.32	3
9	.04	0	1.43	2
10	.02	0	.89	7
11	.01	0	.56	1
12	.00	0	.35	1
13			.22	1
14			.14	0
15			.09	0
16			.06	0
17			.03	0
18			.02	0
19			.01	0
20			.01	0

ported by Pratt (1947) for tests of grouping in the results of Burke Smith's near-liminal visual tests with ESP cards.

In the ESP series, however, we again find a strong concentration of the grouping effect in the higher-scoring runs. Dividing the data into 23 low-scoring runs (2–5 hits, $\bar{x} = 3.83$) and 27 high-scoring runs (6–16 hits, $\bar{x} = 8.22$) produces a W-W Z-score of +1.23 for the lows and −4.31 for the highs ($p < .0001$). The difference between these results is also highly significant ($Z_{diff} = 3.92$, $p < .0001$). Correspondingly, the ANOVA for high runs vs. low runs yields ${}_1F_{48} = 12.3$, $p < .001$, and the Spearman correlation for 50 runs is ${}^r_s = -.394$, $t_{48} = -2.97$, $p < .01$, two-tailed. All 13 of the runs with scores of eight or above show negative Z-scores for grouping, and the total deficit in observed strings of hits and misses, d, amounts to just under 15% for all 27 high-scoring runs together.

In terms of hit-rates, the high-scoring portion of the ESP series overlaps substantially with the low-scoring portion of the visual series. Since these are precisely the places where the grouping effects appear strongest in the two series, respectively, the question arises anew whether the ostensibly visual grouping effect might not indeed be a further expression of Sean's *ESP* grouping effect (there being no serious possibility for the operation of *visual* mechanisms in the ESP task). This seems to me unlikely, for several reasons: First, the conditions of the visual task appear *a priori* likely to engage visual processes rather than psi processes. Second, the overall scoring rate in the visual task is more than twice that of the ESP task, suggesting that the main effects are in fact perceptual in origin. Moreover, the grouping effect in the visual data appears to extend across the whole range of scores, and if it is a visual effect anywhere it seems likely to be a visual effect everywhere. Finally, the grouping effect in that part of the visual data where the overlap in scoring rates is most pronounced (scores of 14 to 22) is actually somewhat attenuated relative to the effect in the remaining low-scoring runs.

Bill Delmore. Overall grouping results for B.D.'s two series involving the 4-choice Schmidt machine are presented in Table 6. For the Kelly/Kanthamani series the main analysis takes the data session-by-session, individual sessions consisting of widely varying numbers of trials. For comparison purposes, Table 6 includes the Z-scores which result for sessions when they are regarded as made up of 100-trial runs; this was exactly the situation for two sessions, approximately the situation for four others, and completely arbitrary for the remaining two. The results are generally consistent with those for the main analysis, and no more will be said about them. The Child series has been analyzed run-by-run as usual.

In neither series is there clear overall evidence of grouping. For

Table 6

GROUPING RESULTS FOR BILL DELMORE (SCHMIDT MACHINE SERIES)

A. Wald-Wolfowitz Tests

1. Kelly & Kanthamani Series (March, 1972)

Session	# Trials	d	E[d]	Dev	% Dev	σ^2 (d)	Z_1	Z_2^a
1	1103	423	437.0	−14.0	−3.2%	172.0	−1.02	−1.04
2	743	292	294.3	−2.3	−0.8%	115.5	−.16	−.40
3	1078	434	441.0	−7.0	−1.6%	179.4	−.49	−.59
4	501	201	201.9	−0.9	−0.4%	80.3	−.04	−.35
5	550	222	231.2	−9.2	−3.9%	96.1	−.89	−.80
6	450	195	188.4	+6.6	+3.5%	77.8	+.69	+.71
7	500	204	211.8	−7.8	−3.7%	88.6	−.78	−.66
8	500	206	214.1	−8.1	−3.8%	90.6	−.80	−.75
Totals	5425	2177	2219.7	−42.7	−1.9%	900.3	−1.41	

2. Child Series (May, 1972)

# Trials	d	E[d]	Dev	% Dev	σ^2 (d)	Z
1800	728	733.3	−5.3	−.7%	282.8	−.29

B. Hit-Group Data

	Kelly/Kanthamani Series		Child Series	
Length of Group	Expected #	Observed #	Expected #	Observed #
1	794.16	772	262.66	253
2	225.88	215	72.34	77
3	64.06	77	19.98	24
4	18.13	16	5.54	3
5	5.11	8	1.55	2
6	1.44	2	.43	0
7	.41	0	.12	0
8	.12	0	.03	0

[a]Z_2 is the Z-score obtained when the session is divided into 100-trial runs.

the Kelly/Kanthamani series, however, there is a fairly strong trend in that direction and seven of the eight sessions show negative Z-scores. The scoring rate was above chance for all sessions and increased almost monotonically across the series; for the first four sessions there is a deficit in strings of −1.75% ($Z = -1.01$), while for the last four the deficit is −3.37% ($Z = -1.49$). For the Child series, division of the data into nine runs with scores of 21 to 27 (average, 24.0) and nine runs with scores of 28 to 37 (average, 31.7) yields for the low scores a 1.6% excess of runs ($Z = +.44$) and for the high scores a 2.8% deficit ($Z = -.87$). It is also striking that among these 18 runs, two show large Z-scores for grouping—a low-scoring run with $Z = +2.53$, and a high-scoring run with $Z = -2.28$.

Particularly in light of the results to be presented subsequently for B.D.'s work with playing cards, I am inclined to regard these

nonsignificant trends in the Schmidt-machine data as dilute expressions of a pattern which is general for B.D. and appears more strongly where his psi is more strongly in evidence. Although his hit-rates for these Schmidt-machine series are statistically highly significant, it should be noted that the *magnitude* of the overall effects is rather small, amounting to a less than 20% excess over MCE and PQ values of around 5–7. By contrast, in the main playing-card experiments he produced hits at 2–4 times MCE with PQ values on the order of 20 to 60.

Analysis of B.D.'s playing-card data presents problems and opportunities partly analogous to those we encountered earlier in conjunction with the Brugmans experiment. Like the "checkerboard" targets, playing cards can be regarded as composed of two attributes. In this case, however, "distance" of a response from its target is not well defined because, although the number attribute can plausibly be regarded as ordered (Ace through King), the suit attribute cannot.

Nevertheless, we certainly have an intuitive sense of proximity that we would like to bring into the analysis in some way. A partial solution is available in the form of the scoring system devised by Fisher (1924), who analyzed responses into nine mutually exclusive and exhaustive classes ordered in terms of increasing proximity to their targets, and associated with correspondingly ordered scores. To utilize this system in the context of the W-W method, I elected to analyze the data repeatedly, each time dividing responses into "hits" and "misses" according to a progressively weaker Fisher-score criterion. The particular score-classes used include all those with scores above the system MCE of zero; in ascending order these are Color-Rank hits (C-R, score = 1.91), Suit-Only (S-O, 4.86), Suit-Rank (S-R, 9.94), Number-Only (N-O, 18.50), Color-Number (C-N, 26.53), and Suit-Number (S-N, 34.55). Thus the first grouping analysis looks only at exact (S-N) hits, the second looks at exact hits plus color-number hits, and so on. This procedure, of course, leads to substantial correlations among the successive W-W grouping analyses since they utilize overlapping information. On the other hand, this overlapping scoring procedure also minimizes the effects of the extreme-value problem described earlier (since we rapidly attain adequate numbers of hits), and the effects of the correlations can be evaluated in the course of the analysis.

By generalizing the scoring for the grouping effect in this way, I hoped to capture evidence of possible shifts in its patterning at different levels of overall scoring. We already know, of course, that in the single-card clairvoyance series and the shuffles series the strong psi-hitting effects are located principally at N-O$^+$ and S-N,

respectively (the superscript "+" means "or higher"; thus, e.g., $N\text{-}O^+ = N\text{-}O + C\text{-}N + S\text{-}N$). Thus, one could reasonably expect that if the data contained any grouping effects they would appear at these levels. Nonetheless, it also seemed possible that if low scoring in a run meant scoring which was systematic but tended to involve the weaker categories of hits (C-R, S-O, S-R), then grouping effects might appear at these lower levels in the low-scoring runs, thus creating systematic differences between low and high runs in terms of where in the vector of grouping scores the effects would show up. Similarly, I hoped that the generalized scoring would help to detect systematic differences related to type of task. To provide a summary index of psi-hitting in the run I also used the Fisher system in the normal way to generate an overall Z-score, taking into account all categories of scoring simultaneously. Thus, each run was represented for purposes of further analysis by its overall Fisher Z-score plus the W-W Z-score for grouping associated with each of the six dichotomization criteria. Analysis then proceeded along the usual lines. To take account of the multiple (and correlated) measures of grouping effects available for each run, extensive use was made of the multivariate counterpart of the familiar univariate ANOVA.

For the 73-run series with Irvin Child, the overall psi effects were very marginal and there was no evidence whatsoever of grouping. Furthermore, exploratory analysis of a few selected runs did not suggest the presence of any internal effects related to scoring rate. I therefore did not carry out the (laborious) further analyses reported for the remaining series.

Overall grouping results for the 46-run single-card clairvoyance series are presented in Table 7. In Part A, both the W-W results and univariate 1-sample ANOVAs based upon the run-by-run grouping Z-scores are displayed for each dichotomization level in turn. The two analyses agree in indicating that there are overall grouping effects, but that these are contributed principally by levels $N\text{-}O^+$ and $C\text{-}N^+$. In the ANOVA data the grouping effect appears to spread more into adjacent categories. This may reflect a genuine consistency, across runs, of effects too small to accumulate to significance by the W-W test. However, we should be cautious here because of the psychometric problems with these Z-scores, particularly at the level of exact hits (S-N). In any event, that the effect is fundamentally confined to the $N\text{-}O^+$ and $C\text{-}N^+$ levels, with spreading to lower levels occurring mainly by way of shared information, is further underscored by the results of the multivariate analysis. As one would expect from the way the six grouping scores are constructed, their correlation matrix takes a regular form in which each measure is highly correlated (typically .6 to .8), with

Table 7

GROUPING RESULTS FOR BILL DELMORE (SINGLE-CARD CLAIRVOYANCE)

A. Grouping Tests

		Wald-Wolfowitz					ANOVA	
Level	d	E[d]	% Dev	σ^2 (d)	Z	p	${}_1F_{45}$	p
S-N	292	297.4	−1.68%	34.8	−.83	n.s.	3.52	.067
C-N$^+$	418	445.8	−6.28%	85.2	−2.96	.005	8.65	.005
N-O$^+$	586	623.7	−6.09%	151.8	−3.02	.003	7.85	.007
S-R$^+$	929	964.0	−3.63%	353.1	−1.83	.07	4.07	.050
S-O$^+$	1067	1100.7	−3.09%	458.3	−1.55	n.s.	3.49	.068
C-R$^+$	1184	1198.8	−1.25%	544.8	−.61	n.s.	.64	n.s.

B. Hit-Group Data

	N-O$^+$		C-N$^+$	
Length of Group	Expected #	Observed #	Expected #	Observed #
1	247.57	215	178.42	154
2	38.52	45	21.22	25
3	7.18	9	3.63	5
4	1.69	3	.85	3
5	.49	0	.24	0
6	.16	1	.07	1
7	.06	1	.02	0
8	.02	0	.01	0
9	.01	0		

adjacent measures and progressively less correlated with those at increasing remove. Because of the high degree of redundancy in the measures, the overall multivariate test using all six measures simultaneously is nonsignificant (${}_6F_{40} = 1.52$, $p < .197$). When the analysis is confined to the three highest-level categories (which is justifiable in light of the previously established pattern of psi scoring in the experiment), the overall test is significant (${}_3F_{43} = 3.10$, $p < .036$). If the psychometrically suspect S-N category is eliminated, the result is ${}_2F_{44} = 4.64$, $p < .015$; but even here it is clear that we are looking at essentially one effect, because the joint significance is sharply reduced from the univariate levels. N-O$^+$ and C-N$^+$ are also correlated .74, and the discriminant analysis produced as a by-product of the multivariate test always assigns them comparable weights.

Accordingly, Part B of Table 7 gives the hit-group data only for the N-O$^+$ and C-N$^+$ analyses. As expected from the negative sign of the overall W-W Z-scores, the effect takes the form of a deficit of singletons compensated by an excess of groups of greater length.

For exploration of possible relationships between grouping effects and overall scoring levels, I next divided the 46 runs into

high-scoring and low-scoring groups using the Fisher Z-scores for the runs as the measure of overall scoring. The 23 low runs had $Z = 0$, SD = .67, while the 23 high runs had $Z = 3.17$, SD = 1.89. These groups were then analyzed for the high-low contrast, again using both the W-W and MANOVA approaches. Looking now particularly at $N\text{-}O^+$ and $C\text{-}N^+$, in the high runs both were significant: $Z = -2.65$ with a 7.4% deficit and $Z = -2.27$ with 6.6% deficit, respectively. The corresponding results for the low runs were $Z = -1.36$ with a 3.8% deficit, and $Z = -2.08$ with a 5.5% deficit. Thus it appears that the tendency toward grouping occurred independently in both sets of runs, but perhaps somewhat more strongly, though not significantly so, in the high-scoring runs.

Parallel indications emerged from the multivariate analyses. First, MANOVA for high vs. low was completely insignificant both overall and for all six dichotomization criteria individually. On the other hand, it is striking that the overall run-score correlates negatively with grouping Z-scores for all criteria above $C\text{-}R^+$, suggesting a tendency for high run-scores to go with more grouping. For $N\text{-}O^+$ in particular, the correlation reaches $-.33$, which approaches significance. If run-score is treated as a covariate in the overall multivariate analysis for grouping (thus eliminating the effects of its correlation with the six criteria), the evidence of grouping presented earlier is largely destroyed, particularly at $N\text{-}O^+$. So again we find suggestions of a weak but possibly systematic relation between overall scoring and grouping.

This lengthy discussion of the single-card clairvoyance data will stand us in good stead as we now turn to the remaining series, for which essentially identical procedures produced generally parallel results.

Overall grouping results for the 55-run shuffles series are presented in Table 8. Again the two forms of analysis are in good agreement, this time suggesting a marginally significant grouping effect confined largely to exact hits, possibly in conjunction with color-number hits. It will be recalled that in this task, as contrasted with the single-card clairvoyance series, the direct evidence for psi resulted almost exclusively from a massive excess of exact hits (over four per run, sufficient to strengthen considerably the psychometric foundations of the S-N analysis). Thus, in both series we have evidence of grouping effects occurring at the levels at which the primary hitting is focused.[6]

[6] It should be noted here that the last two subseries (runs 46–55) of the shuffles experiment were carried out under conditions that were substantially less well controlled than those of runs 1–45, making it appropriate to wonder what effect these later runs might have had on the results. Their exclusion weakens the $C\text{-}N^+$ result ($Z = -1.14$, -2.3%), but slightly strengthens the main result on exact hits ($Z = -2.19$, -4.3%).

Table 8

GROUPING RESULTS FOR BILL DELMORE (SHUFFLES)

A. Grouping Tests

		Wald-Wolfowitz					ANOVA	
Level	d	E[d]	% Dev	σ^2 (d)	Z	p	$_1F_{45}$	p
S-N	430	447.6	−3.93%	65.8	−2.11	.035	5.034	.029
C-N$^+$	514	532.9	−3.55%	91.8	−1.92	.055	4.201	.045
N-O$^+$	707	721.5	−2.01%	159.8	−1.10	n.s.	1.983	.165
S-R$^+$	1148	1172.4	−2.08%	432.4	−1.15	n.s.	1.302	n.s.
S-O$^+$	1331	1347.3	−1.21%	575.6	−.66	n.s.	.111	n.s.
C-R$^+$	1458	1443.2	+1.03%	663.6	+.55	n.s.	.256	n.s.

B. Hit-Group Data

	C-N$^+$		S-N	
Length of Group	Expected #	Observed #	Expected #	Observed #
1	218.04	200	182.40	168
2	22.56	30	15.90	22
3	2.93	5	1.78	3
4	.53	0	.28	0
5	.13	0	.05	0
6	.03	0	.01	0
7	.01	0	.00	0

Part B of Table 8 supplies the hit-string data for the C-N$^+$ and S-N analyses. As previously, the effect takes the form of a deficit of singletons and an excess of groups of greater length.

The 55 runs were then divided, using as before the Fisher run-scores, into 26 high-scoring and 29 low-scoring runs. MANOVA produced no evidence of systematic differences in grouping between these scoring levels, either overall or with any one of the six dichotomization criteria. However, use of the Fisher Z-score as a covariate again destroys the overall evidence of grouping at C-N$^+$ and S-N, suggesting a possible weak relationship between run-scores and grouping. The W-W analysis likewise indicates a tendency toward grouping in both sets of runs, nonsignificantly stronger in the high-scoring runs.

It should be pointed out, however, that the high/low contrast here was particularly poor, since the "low-scoring" runs themselves contained highly significant evidence of psi ($Z = 5.4$ vs. $Z = 8.06$ for the high-scoring runs). It also occurred to me at this point that because the grouping effects follow so closely the patterns of the primary psi-hitting, analysis for possible relations between the two might be sharpened if the measure of hitting were based specifically on the relevant scoring categories rather than on all categories simultaneously, as in the Fisher analysis. This idea is most readily pursued in the context of the shuffles data, where the number of

exact hits provides an index of overall scoring. Thus, the 55 runs were again divided, this time into 27 lows (0–3 exacts, 2.04 per run) and 28 highs (4 and up, 5.93 per run), and the W-W analysis recomputed. For S-N this leads to $Z = 0$ in the low runs and -2.27 in the high runs, while for C-N$^+$ the corresponding numbers are $+.87$ and -2.58. Thus it seems that my attempt to gain generality through use of the Fisher Z-score for the main analyses may have led to some insensitivity to the tendency for grouping effects of very specific form to correlate with overall scoring levels in B.D.'s data.

A related point is that, for reasons analogous to those we encountered in the Brugmans/Van Dam results, the methods used here very likely underestimate substantially the amount of grouping in B.D.'s psi performances, in part by failing to take sufficiently detailed account of proximity relations between responses and their targets. For example, a triad consisting of an ace-of-spades (AS) call to a two-of-clubs (2C) target sandwiched between two exact hits is scored essentially as two isolated hits, although we know because of B.D.'s imagery-based method of response (Kelly, Kanthamani, Child, and Young, 1975) that he was very likely close to getting a third exact hit. Likewise, trials such as QH/KH and 6S/8S occurring adjacent to exact hits are regarded only as suit (or suit-and-rank) hits, when in fact he was much closer. These three examples and several others all come from a single 25-trial stretch without feedback in run 35 of the single-card clairvoyance series.[7]

The last series to be considered is the series of 75 visual runs, for which overall grouping results are presented in abbreviated form in Table 9. Note that in this series the hit rate is in excess of 50%, so that we are typically working at the opposite side of the symmetrical distribution for d. Thus the extreme-value problem in this series is most acute for the C-R$^+$ analysis (since there are hardly any

---

[7] A related methodological point can be made here which applies to all series. Consider the following two hypothetical sequences, each containing six hits (1's) and 19 misses (0's):

(1) [1110 . . . 01110]
(2) [111011100 . . .]

The second of these sequences looks far more interesting as regards grouping of hits; yet both sequences contain four strings of hits and misses, and thus they are equivalent from the point of view of the Wald-Wolfowitz analysis. That is, the methods used here are completely insensitive to relations of proximity weaker than strict adjacency between hits. My informal impression is that a statistical technique capable of utilizing such proximity information would generally further strengthen the results reported here. In short, the W-W test, while certainly useful, is not very powerful.

Table 9

GROUPING RESULTS FOR BILL DELMORE (VISUAL SERIES)

	Wald-Wolfowitz						ANOVA	
Level	d	E[d]	% Dev	σ^2 (d)	Z	p	$_1F_{74}$	p
S-N	1512	1557	−2.89%	590.1	−1.84	.066	3.386	.070
C-N$^+$	1452	1500	−3.20%	556.9	−2.01	.045	4.335	.041
N-O$^+$	1333	1378	−3.27%	487.9	−2.00	.046	3.792	.055
S-R$^+$	784	798	−1.75%	224.0	−.93	n.s.	2.095	.152
S-O$^+$	716	720	−.56%	184.0	−.24	n.s.	1.194	n.s.
C-R$^+$	478	482	−.83%	93.6	−.35	n.s.	2.400	.126

misses), and the numbers of observed strings, etc., *decrease* as we proceed from S-N to C-R.

The results again suggest a weak tendency toward grouping of hits, concentrated at the level of number hits and above. The hit-group tables are too lengthy to include here (because of the high hit rate), but follow a pattern of deficits at short lengths, compensated by excesses erratically distributed over greater lengths. No relationships were discovered for this series between scoring and grouping.

Finally, MANOVA analyses were carried out to make exploratory statistical comparisons among B.D.'s three main playing-card series in terms of the overall patterns of grouping effects defined by the six dichotomization criteria as applied to their constituent runs. Not surprisingly, in view of the similarity in form of the effects described above, no significant differences of any kind were found between the single-card clairvoyance series and the shuffles series, or between the pooled psi series and the visual series.

DISCUSSION

The results presented so far show clearly that nonrandom distribution of hits within runs is a feature sometimes associated with strong psi performance. We have next to consider the sources and interpretation of these effects.

I will first comment briefly on the results of the visual series. Pratt (1947) had already pointed out that the grouping effects arising in "near-liminal" visual tests were almost certainly due primarily to peripheral factors such as shifts in ambient lighting, retinal adaptation, and so on, possibly interacting with more general and slowly-varying features of psychophysiological state. Hence it would be natural to expect that conditions favorable for correct

perception of the stimulus, once established, would tend to persist for a period of time extending over a sequence of closely-spaced trials. The *psi* data available to Pratt, however, afforded no compelling evidence of nonrandom grouping, and thus suggested to him a fundamental contrast in this respect between psi and perceptual processes. Pratt interpreted this contrast tentatively as due to the dependence of success in psi tasks upon largely unconscious and rapidly changing central events. I believe Pratt was correct in regarding success in psi tasks as relatively dependent upon central factors. However, the data assembled in the present report show clearly that the *factual* contrast he proposed—i.e., presence vs. absence of grouping effects in perceptual vs. psi tasks—is not fundamental. Indeed, the main value of that comparison now seems to me to lie rather in providing examples, for two subjects, of psi effects that compare favorably in magnitude with their perceptual counterparts.

Turning now to possible factors involved in the production of grouping effects in psi data, we should begin by considering two relatively uninteresting "artifactual" sources. The first is simply failure of conditions, providing opportunities for sensory leakage, cheating by subjects or experimenters, etc. The four series most suspect in this regard, to my mind, are the Harribance series, the shuffles series with B.D., the Van Dam series, and the Soal-Stewart series. The Harribance data, as indicated, were deliberately collected under moderately informal conditions. However, I see nothing in the description of these conditions (Morris, 1972) that would lead one to expect the kind of intermittent brief failure that would be required to produce the reported results. In the shuffles series with B.D., the last two subseries were definitely (and deliberately) methodologically weakened. However, the pattern of grouping effects contained within these weaker runs appears similar overall to that of the remaining runs, and their exclusion does not destroy the main result. Additional reasons for taking seriously even the weaker portions of the shuffles data are provided in Kanthamani and Kelly (1974b, 1975). In the Van Dam series, the possibility of auditory leakage cannot be absolutely ruled out, and if auditory leakage occurred it might be expected to lead to grouping effects, for reasons parallel to those adduced above in discussing the visual data. However, Schouten and Kelly (1978, p. 285–288) have marshalled arguments which seem to me to render this explanation in terms of auditory cues highly unlikely. The Soal-Stewart series is by all odds the most suspect of the four, and must necessarily remain so pending resolution of the present controversy regarding Soal's "reliability." However, in the absence of an empirically supported cheating hypothesis which could also

account for the form of the grouping effects reported here, I am inclined to regard these effects also as genuine.

The second possible artifactual source to consider is *position effects*—nonrandomness in the aggregate distribution of scoring with respect to location in runs, sessions, or series. Although it is clear that grouping effects *can* sometimes be produced as a secondary consequence of position effects, position effects can be largely excluded as a source of the grouping effects reported here. For the Van Dam and Stewart data, run and series effects known to exist have been explicitly minimized by the strategies of analysis used. Of the remaining series, the only one that to my knowledge displays a strong position effect is the Delmore 4-choice Schmidt machine PK series with Irvin Child which contains a significant quadratic trend (terminal salience) across fifths of the run, but only a suggestion of a grouping effect. Inspection of individual runs containing strong trends toward grouping, moreover, suggests that the long strings of hits have no particular locational preference. As was already noted by Pratt (1947), run-position effects and grouping effects may well be psychologically independent phenomena, each capable of occurring or not occurring in conjunction with the other. For the future, it would be possible to investigate such relationships analytically, for example by scoring each run separately for linear and quadratic trends and for grouping, and correlating these scores across the series.

For the present, however, we are left with a reduced number of plausible candidates to consider as possible factors in the production of the observed grouping effects. Factors that come immediately to mind include type of task, trial rate, feedback regime, theoretical hit-probability, experimenters, and subjects. Some data pertinent to assessing these possibilities are summarized in Table 10.

Clearly the set of contrasts available here is imperfect in many ways, primarily because we must rely upon post-hoc comparisons rather than experimental ones, using data originally collected for entirely extraneous reasons. Thus, many of the contrasts plainly confound multiple factors, and there could conceivably be additional pertinent factors lurking in conjunction with some of the series, but unrecognized. It would be particularly desirable to have more data involving PK tests (including REG work) and to investigate grouping effects in psi-missing data.

Nevertheless, the information already available is sufficient, I believe, to point the way toward correct interpretation of the overall meaning of the grouping results.

First, it seems clear that the presence of psi, as measured for example by direct-hitting rates, is a necessary though not sufficient

Table 10

SUMMARY OF SOME FEATURES PERTINENT TO INTERPRETATION OF GROUPING EFFECTS

Subject	Series/ Exptr	Type of Task	Direct Psi Scoring	Type
Miss S	Riess	GESP	Extremely strong	Hitting
Pearce	Pearce/Pratt	Clairvoyance	Strong	H
Stewart	Soal/Stewart	GESP	Strong	H
Van Dam	Brugmans, Heymans, & Weinberg	GESP	Extremely strong	H
Harribance	Morris, 1000 runs (Klein, Exp)	Clairvoyance	Fairly strong	H
Delmore	Child	Schmidt machine (precog?)	Moderate	H
Delmore	Kelly	Schmidt machine (PK mode)	Moderate	H
Delmore	Child	Single-card clairvoyance	Weak	H (suits)
Delmore	Kanthamani & Kelly	Single-card clairvoyance	Very strong	H ($N\text{-}O^{+}$)
Delmore	Kanthamani & Kelly	Shuffles (PK?)	Very strong	H (S-N)

condition for the appearance of within-run grouping effects. Necessity is hardly surprising, and is indicated not only by the Child single-card clairvoyance data from the present study (Table 10, Row 8), but also by several other series I have analyzed (including two from Sean Harribance) which show no trace either of psi or of grouping effects. Insufficiency is indicated, for example, by the results from the Riess series and several of the other high-scoring series analyzed by Pratt.

Second, examination of Table 10 strongly suggests that external parameters of the experimental situation such as type of psi task, response time, feedback regime, and p(hit) cannot be primary sources of the grouping phenomenon, although they seem quite likely to influence the form and magnitude of its expression. So far as we can now judge, grouping effects occur with a variety of task types, with feedback at levels ranging at least from the trial to the session, and with a substantial range of values for both p(hit) and time per response. Similarly, a contrast such as that between the Riess and Pearce/Pratt series holds several of these features more

Table 10 (*Continued*)

Nonrandom Grouping Present?	Related to Scoring Level?	Time per Response	Feedback Level	*p*(hit)
N	—	1 min reg	None	1/5
Y	Y	1 min reg	Session (1-3 runs)	1/5
Y	Y	2-3 sec reg	Run	1/5
Y	Y?	Secs to mins, highly variable	Session? (6-47 trials)	1/48 1/6
Y	Y	A few sec	Run	1/5
?	Y?	A few sec	Trial	1/4
Y?	Y?	A few sec	Trial	1/4
N	—	A few sec to a min	Trial	1/4
Y	Y	A few sec to a min	Generally each trial, sometimes 13-26 trials	N-O=1/13 S-N=1/52
Y	Y	Hard to define	Run	C-N=1/26 S-N=1/52

or less constant; yet the latter series displays a strong within-run grouping effect while the former does not.

The two most promising candidates appear rather to be the *experimenter* and the *subject*. That the experimenter, although again quite possibly conditioning the expression of the grouping phenomenon, is not its primary source would be strongly indicated by two types of occurrence: (a) cases in which different subjects produce different grouping results with the same experimenter; and (b) cases in which a given subject produces comparable grouping effects in conjunction with different experimenters (other things remaining equal, of course, in both cases). A good example of Type 1, if genuine, is provided by the contrast between the Soal/Shackleton and Soal/Stewart grouping results. Moreover, the Van Dam series presents a good example of Type 2, since this subject produced grouping effects spanning all three of his experimenter/agents: For example, the group of eight hits in Session 2 included three trials with Brugmans as experimenter, three with Heymans, and three with Weinberg; the group of 12 hits in Session 3 involved

Brugmans and Weinberg alternately; and the group of six in Session 12 involved just Heymans, Thus, although more data pertinent to this contrast are sorely needed, what evidence we have at present tends to point toward the *subject* as the primary source of the grouping effect.

Here we arrive at the central point of this paper: The grouping effects reported for these exceptional subjects seem to me best interpreted as indications of sporadic and transient establishment, within them, of conditions unusually conducive to successful psi response. This interpretation, furthermore, does not depend exclusively on negative evidence regarding the experimenter; I will now briefly summarize a variety of circumstantial evidence suggesting not only that specially psi-conducive conditions are occurring in the subjects, but also that these conditions may in some cases have measurable physiological concomitants, and that they may even on occasion be recognized by the subject as an alteration of conscious state. (In none of this, incidentally, do I mean to imply that the relevant conditions are necessarily, or even likely, the *same* for all subjects.)

For Hubert Pearce and Gloria Stewart we unfortunately have essentially no information, neither Rhine nor Soal apparently being much given to this sort of inquiry.

In the case of Van Dam, however, we have the remarkable report by Brugmans (1924) which—anticipating the recent emergence of "converging-operations" research—provides information both about Van Dam's subjective experiences *and* about concomitant behavioral and physiological events. Van Dam reported introspective awareness of three aspects of his internal state associated with success: First and most important, the best results occurred in conjunction with what he called his "passive state," a voluntarily induced, subjectively distinct state that appears to have been characterized in part by deep relaxation and mental quietude. Two less distinct sensations more variably associated with success were a feeling of "contact" with the experimenters, and a feeling of having successfully completed the task. Brugmans points out in passing that Van Dam's overall level of success (60 hits in 187 trials, up to that point) shows that even the latter two sensations in fact corresponded, though imperfectly, to something objective in the situation. The bulk of his report, however, is concerned with a study of physiological changes associated with Van Dam's "passive state." Three kinds of measures were actually taken—pulse, respiration, and a kind of GSR recorded between cylindrical electrodes taped into Van Dam's palms—but only the electrodermal results are discussed in the report. Although Brugmans does not supply

the kind of quantitative detail that we would expect nowadays, he illustrates his remarks with over a dozen photographs of the original physiological tracings and is careful to note that these illustrations are representative of the (unspecified) remainder. He states that Van Dam's transition to the passive state was reliably marked by changes in the record analogous to changes occurring with sleep onset. Electrodermal response to external stimuli was also markedly enhanced during "passivity," and the combination of tonic levels and responsiveness further served to distinguish the genuine passive state from one of normal relaxed wakefulness. Van Dam's verbal state-reports and these physiological patterns were highly consistent, according to Brugmans. It is unfortunate that the complete original records are no longer available for more detailed analysis, particularly since visual inspection of the available photographs indicates that systematic changes were also occurring in the other physiological channels—changes which might have helped to illuminate further the nature of the passive state. On the other hand, in light of results from Schouten and Kelly (1978), we can now add one further detail: that hitting trials in general were also *behaviorally* distinct from missing trials, with markedly shorter average response times. In sum, taking all sources of evidence into account it appears reasonable to suggest that in this experiment Van Dam's successful responses in general, and his long bursts of hits in particular, arose essentially as motor automatisms out of a background supplied by a transient, mildly altered state of consciousness.

In the case of Sean Harribance we have the important paper of Morris et al. (1972) showing that in each of two highly successful ESP series high-scoring runs, in comparison to chance runs, were marked by significantly greater abundance of alpha rhythms in Sean's occipital EEG. Interpreting this EEG result with unusual thoroughness and care, these authors concluded that it was most probably genuinely reflective of significant differences in Sean's internal state, the high-scoring runs being associated with a relatively relaxed state of passive, inwardly-directed awareness. It is unfortunate that the ESP records from these *specific* experiments were not available for grouping analysis (the records are in storage along with other materials belonging to the recently-moved Psychical Research Foundation, and temporarily inaccessible). However, the records I did analyze come from a period following the second experiment by just a few months. Therefore it seems reasonable to suppose that similar grouping effects probably characterized Sean's strong ESP performance during the earlier EEG-ESP work. The scoring rates, certainly, are very similar. If this is true,

then we also know that a second striking difference between the strong and weak ESP runs consisted in a tendency for unusually long bursts of consecutive hits to be injected sporadically into the high-scoring runs. One wonders whether the EEG-ESP relationship unearthed by Morris et al. (1972) might not be a diluted reflection, at the level of the entire run, of a relationship which would prove even stronger at a more molecular level; that is, whether the surplus of alpha rhythms in high-scoring runs might not be associated primarily with these occasional long strings of hits. It would also be interesting to determine whether this EEG-ESP relation works in the opposite direction as well; that is, whether dividing runs or trials on the basis of EEG criteria would lead to significant differences in ESP scores between the resulting groups. If so, the EEG criteria would evidently be associated with sufficient as well as necessary conditions for Sean's strong scoring (see Stanford and Palmer, 1975).

For Bill Delmore we unfortunately have no useful physiological data. However, several observations point to the sporadic occurrence in him of unusually psi-conducive conditions associated both with bursts of hits and with awareness of success. For example, a conspicuous feature of his informal demonstrations is their streakiness; "hot" periods in which everything worked were often flanked by periods in which nothing much of interest happened. Like many other features of his spontaneous performances, this tendency was also reflected to a limited degree in his experimental work. In the single-card clairvoyance series, for example, there were a few occasions on which he felt these hot streaks coming on and requested a block of trials without feedback. The results for these episodes were often spectacular: For example, a block of 13 trials at the beginning of run 21 contains nine number hits, including a string of seven, and three near-misses (QC/JC, 5D/4D, QS/JS). Another block of 13 trials without feedback at the end of this same run contains six exact hits and five more near-misses. A third example is the 25-trial group in run 35 mentioned previously, containing seven exact hits and at least 10 near-misses. In this last episode the subjective awareness of success was particularly conspicuous, since all seven exact hits were also confidence calls. There was one additional confidence call in the group, and it produced a color-number hit. It should be added here, however, that B.D.'s awareness of success was far from complete, since both the shuffles series and the single-card clairvoyance series contained numbers of unrecognized hits far in excess of MCE. Nevertheless, it is clear that sporadic episodes of unusually high scoring occurred in B.D., and that on at least some occasions these were associated with experienced changes in his subjective state. Whether these

changes would have been reflected in physiological measures we have no way of knowing.[8]

For a final example I will draw upon the work of G.N.M. Tyrrell with Gertrude Johnson, using the hidden-light task he developed specifically for her. The fully-developed apparatus included facilities for automatically recording on a strip-chart the sequence of hits and misses. I quote at length Tyrrell's (1938) interesting comments on these records:

> The tape records, in cases where the score has been well above the chance-expected 20 per cent., are very instructive. They show that the increased score is due to short bursts of successes interposed here and there on the chance-successes. There may be groups of from six to ten consecutive successes in one or two places in an otherwise chance record of a hundred trials. G.J. is quite definitely aware, during the periods when these groups occur, of almost losing consciousness of her surroundings. She says that a peculiar, and rather exalted feeling comes over her, making her feel that it is almost impossible to fail, and so long as this lasts, the successes follow one another in an almost unbroken chain; but it is never maintained for more than a few seconds at a time. One sees here a kind of mental dissociation (it may be not unlike that which often accompanies automatic writing) which lets the extra-sensory faculty overcome its customary inhibitions for a moment or two (p. 108).

Although I have not been able to obtain copies of Tyrrell's tape records, a moderate amount of calculation suffices to show that even the weakest of the events he describes—viz, single groups of six hits injected into random sequences of a hundred trials—not only would account for the observed scoring rate but would also rapidly be detected by the Wald-Wolfowitz grouping test over a short series of runs.

---

[8] Irvin Child has pointed out in correspondence that for the grouping effect in B.D.'s shuffles data, the straightforward "state" interpretation may be somewhat strained given the nature of the task. That is, if we assume that psi occurred during the shuffling, then grouping of hits seems to imply very strong PK effects operating over short stretches of time—perhaps only small fractions of seconds. I do not find this implausible, particularly since the unusually strong psi bursts might arise out of a psi-conducive background state of longer duration, and their effects could accumulate over a series of shuffles. For further investigation, it would be particularly desirable to keep track of all of the intermediate configurations of the shuffled deck, as well as its final configuration, so that we could study in detail the manner in which the results converge toward their final form. It would also be germane to look for pulsed PK effects on fast random event generators. Another interpretation might be that the PK effects actually occurred retroactively during the card-by-card checkup following the shuffling. Although this interpretation strikes me as *a priori* considerably less palatable, it does have the merit of removing the apparent asymmetry in task structure between the shuffles series and the other series. The two interpretations, moreover, lead to differential expectations as to where (if anywhere) in the course of the task physiological or other predictors of success might be found.

Certainly, none of these individual pieces of testimony are without serious flaws. Most importantly, in every case the amount of physiological and phenomenological detail retrospectively available is far less than we would like. Nevertheless, the examples are consistent in pointing to the phenomenon of grouping of hits as a potentially vital nexus in the search for psychophysiological conditions associated with unusually strong psi performance.

Finally, the practical significance of these grouping effects becomes even clearer if we now briefly consider the *magnitude* of their contribution to the overall evidence for psi in the series we have analyzed. This is not a straightforward calculation, but we can at least roughly bracket the range with approximate estimates of upper and lower bounds. To do this we first calculate the total deviation from MCE. Now, ignoring random variation, we want to know what proportion of this total excess is contributed by strings of unusual length. To obtain a lower bound we can take as our numerator the excess of observed over expected hits beyond the crossover point in hit-group lengths, using expectations calculated in the normal way on the basis of the observed hits, as in Tables 1–9. Note, however, that this approach is conservative because if we think of psi as sporadically injecting excess hits into an otherwise largely random sequence, inclusion of these excess hits in the calculations has already artificially inflated the "chance" expectations for the longer strings. For an *upper* bound, I therefore adopted a model in which these expectations were instead calculated on the assumption of totally random scoring. The proportion of the total excess of hits contained in the long groups of hits was then calculated as before. I have made these exploratory calculations for three series, with impressive results; For the B.D. single-card series, the proportion of $N\text{-}O^+$ hits accounted for by long strings lies between 18 and 70%. For Soal-Stewart Series II, it is between 37 and 87%. And for the Harribance data, it is between 67 and 100%. Although these bounds are soft, and rather wide, it can be concluded with confidence that very sizeable proportions of the overall psi effects in these series are specifically due to grouping effects. Note too that the strong tendency we have observed for grouping effects to congregate in high-scoring runs is consistent with this picture. Tyrrell's at first startling contention that Gertrude Johnson's *entire* psi performance was reducible to sporadic bursts of this sort proves after all to be quite in line with these results for other subjects.

Conclusions

To my mind, systematic internal effects in psi data often testify even more strongly than high scoring rates to the psychological

reality of psi processes. The grouping effects reported here have for me this property because they seem to reveal, for several of our outstanding psi performers, systematic fluctuations in whatever internal conditions were associated with their capacity to generate the phenomena.

In further efforts to identify the *nature* of these conditions, grouping effects will clearly be of great practical value. For the basic strategy of such research is to seek contrasts, in terms of whatever variables are being studied, between sections of the record that contain psi and sections that do not. Grouping effects, where they occur, can help us to identify with much greater confidence sections of the record containing psi, and thus substantially increase the precision of these experimental contrasts.

For example, in our initial physiological work with Sean Harribance (Kelly and Lenz, 1976) we compared all hitting trials with all missing trials in terms of the spectral content of Sean's parietal EEG during the two-second periods preceding his responses. As it happens, this single session contained evidence neither for psi nor for grouping: but had the psi results been like those reported here from Sean's earlier work, it clearly would have been preferable to contrast missing trials not with *all* hitting trials—since these certainly would include numerous chance hits—but only with hits occurring in consecutive groups of four or more. An "isolation" effect such as produced by Pearce could also be useful, since it would evidently permit a contrast, particularly in high-scoring runs, between psi-conducive conditions on isolated hitting trials and conditions on adjacent missing trials that were actively antagonistic, rather than merely indifferent, to the expression of psi.

Clearly such applications are most readily pursued in the context of intensive longitudinal investigations with carefully selected subjects. The example of Pearce's atypical effect underscores again the importance of careful attention to individual differences in such research. Although the results so far suggest a considerable amount of between-subjects variation in the patterning of grouping effects, it remains to be seen how responsive these patterns may be, in given individuals, to variations such as reviewed earlier in the structure of experimental conditions. It might prove possible, for example, to modify the task environment so as to "shape" a subject's grouping effect optimally for whatever experimental purposes are at hand. Alternatively, if grouping patterns prove to be highly individualized and stable, they might be put to work as signatures of personal identity—"mindprints" in Eisenbud's apt phrase—useful in tracing sources of psi effects.

The work reported here focuses on psi-conducive conditions occurring transiently within individual runs, and typically involving

only short sequences of trials. This level of analysis is of course particularly germane to the kinds of detailed psychophysiological investigations outlined above, but it should also be recognized that it provides only a systematically limited view of the subject matter. I believe that by opening our analysis window beyond the individual trial to segments of greater length, we would begin to see additional systematic fluctuations in psi performance having longer time courses and quite possibly related to different sets of factors in the respondents. For example, both Bill Delmore and Van Dam produced spectacular outbursts of hitting coincident with large portions of runs, as did the subject of Musso and Granero (1973). Furthermore, although I have not presented any of these results here, I know that in several of the series analyzed for this report there are also very strong continuities from *run* to *run* in scoring. In a sense virtually the entire 74 runs of the initial Riess experiment can be regarded as a psychological unit, one which might profitably have been studied intensively with respect to more global psychophysiological conditions holding in Miss S during that entire period and possibly associated with her extraordinary scoring.

Finally, I find it particularly exciting to discover, even within these bland forced-choice testing environments, traces of what appears to be an important connection between certain altered states of consciousness and unusually strong psi performance. I have come increasingly to believe that this intersection lies at the heart of our subject matter, and the results of the present investigation if anything have further strengthened this belief. However, as my views on this subject have recently been expressed in considerable detail elsewhere (Kelly and Locke, 1981), I will not attempt to restate them here.

References

BRUGMANS, H. I. F. W. "L'Etat passif" d'une télépathe, controlé par le phénomène psychogalvanique. *L'Etat Actuel des Recherches Psychiques d'après les Travaux du IIème Congrès Internationale*. Paris, 1924, 95–125.

BURDICK, D. S., AND KELLY, E. F. Statistical methods in parapsychological research. In B. B. Wolman (Ed.), *Handbook of Parapsychology*. New York: Van Nostrand Reinhold, 1977.

CADORET, R. J., AND PRATT, J. G. The consistent missing effect in ESP. *Journal of Parapsychology,* 1950, 14, 244–256.

FISHER, R. A. A method of scoring coincidences in tests with playing cards. *Proceedings of the Society for Psychical Research,* 1924, 34, 181–185.

KANTHAMANI, H., AND KELLY, E. F. Card experiments with a special subject. I. Single-card clairvoyance. *Journal of Parapsychology,* 1974, 38, 16–26. (a)

KANTHAMANI, H., AND KELLY, E. F. Awareness of success in an exceptional subject. *Journal of Parapsychology,* 1974, 38, 355–382. (b)

KANTHAMANI, H., AND KELLY, E. F. Card experiments with a special subject. II. The shuffle method. *Journal of Parapsychology,* 1975, 39, 206–221.

KELLY, E. F., CHILD, I. L., AND KANTHAMANI, H. Explorations in consistent missing. *Journal of Parapsychology,* 1974, 38, 230–231.

KELLY, E. F., AND KANTHAMANI, H. A subject's efforts toward voluntary control. *Journal of Parapsychology,* 1972, 36, 185–197.

KELLY, E. F., KANTHAMANI, H., CHILD, I. L., AND YOUNG, F. W. On the relation between visual and ESP confusion structures in an exceptional ESP subject. *Journal of the American Society for Psychical Research,* 1975, 69, 1–31.

KELLY, E. F., AND LENZ, J. E. EEG correlates of trial-by-trial performance in a two-choice clairvoyance task: A preliminary study. In J. D. Morris, W. G. Roll, and R. L. Morris (Eds.), *Research in Parapsychology 1975.* Metuchen, N. J.: Scarecrow Press, 1976.

KELLY, E. F., AND LOCKE, R. G. *Altered States of Consciounness and Psi: An Historical Survey and Research Prospectus.* (Parapsychological Monographs No. 18.) New York: Parapsychology Foundation, 1981.

MARKWICK, B. The Soal-Goldney experiments with Basil Shackleton: New evidence of data manipulation. *Proceedings of the Society for Psychical Research,* 1978, 56, 250–281.

MOOD, A. M., AND GRAYBILL, F. A. *Introduction to the Theory of Statistics.* New York: McGraw-Hill, 1963.

MORRIS, R. L. Guessing habits and ESP. *Proceedings of the Parapsychological Association,* 1972, 8, 72–74.

MORRIS, R. L., ROLL, W. G., KLEIN, J., AND WHEELER, G. EEG patterns and ESP results in forced-choice experiments with Lalsingh Harribance. *Journal of the American Society for Psychical Research,* 1972, 66, 253–268.

MUSSO, J. R., AND GRANERO, M. An ESP drawing experiment with a high-scoring subject. *Journal of Parapsychology,* 1973, 37, 13–36.

MUSSO, J. R., AND GRANERO, M. U-effects in an ESP experiment with concealed drawings. *Journal of Parapsychology,* 1981, 45, 98–120.

PRATT, J. G. Trial-by-trial grouping of success and failure in psi tests. *Journal of Parapsychology,* 1947, 11, 254–268.

PRATT, J. G. Computer studies of the ESP process in card guessing. I. Displacement effects in Mrs. Gloria Stewart's data. *Journal of the American Society for Psychical Research,* 1967, 61, 25–46.

RHINE, J. B., AND PRATT, J. G. A review of the Pearce-Pratt distance series of ESP tests. *Journal of Parapsychology,* 1954, 18, 165–177.

RIESS, B. F. A case of high scores in card guessing at a distance. *Journal of Parapsychology,* 1937, 1, 260–263.

RIESS, B. F. Further data from a case of high scores in card-guessing. *Journal of Parapsychology*, 1939, 3, 79–84.

SCHOUTEN, S. A., AND KELLY, E. F. On the experiment of Brugmans, Heymans, and Weinberg. *European Journal of Parapsychology,* 1978, 2, 247–290.

SIEGEL, S. *Nonparametric Statistics for the Behavioral Sciences.* New York: McGraw-Hill, 1956.

SOAL, S. G., AND BATEMAN, F. *Modern Experiments in Telepathy.* New Haven: Yale University Press, 1954.

SOAL, S. G., AND PRATT, J. G. ESP performance and target sequence. *Journal of Parapsychology,* 1951, 15, 192–215.

STANFORD, R. G., AND PALMER, J. Free-response ESP performance and occipital alpha rhythms. *Journal of the American Society for Psychical Research,* 1975, 69, 235–243.

STEVENS, W. L. Distribution of groups in a sequence of alternatives. *Annals of Eugenics,* 1939, 9, 10–17.

SWED, F., AND EISENHART, C. Tables for testing randomness of grouping in a sequence of alternatives. *Annals of Mathematical Statistics,* 1943, 14, 66–87.

TYRRELL, G. N. M. *Science and Psychical Phenomena.* New York: Harper & Row, 1938. (Reprinted, in the same volume with *Apparitions,* in 1961 by University Books.)

WALD, A., AND WOLFOWITZ, J. On a test whether two samples are from the same population. *Annals of Mathematical Statistics,* 1940, 11, 147–162.

Department of Electrical Engineering
Duke University
Durham, North Carolina 27706

APPENDIX A

1. *Wald-Wolfowitz Test*

The total number of groups, d, of hits and misses in a sequence of N trials containing m hits and n misses follows a known distribution with exact mean and variance:

$$E[d] = \frac{2mn}{m+n} + 1 \tag{1}$$

$$\sigma^2(d) = \frac{2\,mn\,(2mn-m-n)}{(m+n)^2\,(m+n-1)} \tag{2}$$

The distribution of d rapidly approaches normality, and for $m, n > 10$, the approximation

$$Z = \frac{d - E[d] \pm .5}{\sigma(d)} \tag{3}$$

can be used in conjunction with a table of the normal distribution to estimate the probability of observing a number of groups $\leqslant d$ for small d (indicating grouping) or $\geqslant d$ for large d (indicating isolation). To analyze a collection of N-trial runs, d, $E[d]$, and $\sigma^2(d)$ are computed separately for each run and summed over runs as appropriate to provide an overall test using (3). For a given N, the values of $E[d]$ and $\sigma^2(d)$ approach their maxima as $m \to n$, with their values falling off symmetrically around the central point ($N/2$ for odd N, or $(N+1)/2$ for even N). In practice, this means that runs with very few hits (or very many hits) tend to make relatively small contributions to the aggregate results.

II. *Distribution of Lengths for Groups of Hits*

If m and n are again the total number of hits and misses in a sequence of N trials, and m_k is the number of groups in the sequence containing exactly k hits, then

$$E[m_k] = \frac{m^{(k)}}{N^{(k+1)}}[n(n+1)]. \tag{4}$$

Here $m^{(k)}$ is the kth factorial power of m, defined as $m(m\text{-}1)(m\text{-}2) \ldots (m-k+1)$. Likewise, $N^{(k+1)} = N(N-1)(N-2) \ldots (N\text{-}k)$. The expectations are calculated for $k = 1, 2, \ldots, m$.

For a single run,

$$\sum_{k=1}^{m} k \cdot m_k = \sum_{k=1}^{m} k \cdot E[m_k] = m. \tag{5}$$

To characterize the form of grouping effects in a series of runs, observed and expected numbers of hit-strings for all relevant lengths are calculated separately for each run and summed over runs. Since relation (5) holds for each run separately, it also holds in the aggregate.