SIZE OF STIMULUS SYMBOLS IN EXTRA-SENSORY PERCEPTION

J. G. PRATT and J. L. WOODRUFF
Duke University

Abstract: An investigation of four problems is reported: (1) Does ESP occur? (2) If so, what is the relation between level of scoring and size of symbols? (3) What is the effect of experience in formal ESP tests on rate of scoring? (4) What is the relation of "newness" of stimulus material to rate of scoring?

The entire research involved the participation of 66 subjects who were tested to the extent of 3,868 runs of 25 trials each with ESP cards. The total number of hits scored was 970 beyond mean chance expectation, an average of 5.25 hits per run, which gives a critical ratio (C.R.) of 7.80. These are total results from two series which are distinguished on the basis of differences in experimental conditions.

In one of these, Series B, two experimenters were present at every test and certain special safeguards against error were used here for the first time. In this series, 32 subjects made 2,400 runs with a positive deviation of 489 hits. The C.R. is 4.99.

No significant differences in scoring rates are found in relation to symbol sizes in the experiment as a whole. No direct relation is found between the experience of subjects and the rate of scoring.

The use of "new" material is found to give scores which are significantly higher than those obtained with "old" material. "Experienced" subjects scored as well with "new" material as "inexperienced" subjects. A decline in the effectiveness of "new" material with successive sessions of its use is noted. The advantage in favor of higher scores with "new" material was greater when the subjects knew what symbol size was being used.

INTRODUCTION

Background of the Research

Any research based on the hypothesis of ESP involves, in a sense, the problem of the re-testing of that hypothesis. Without adequate evidence that the phenomenon itself is present, any problem concerned with the nature of ESP has little chance of solution. In such a sense, this research was again a test of the primary hypothesis. At the same time, however, the major goal of the research at the point of its inception was to ascertain whether there is any relation between the size of symbols used as stimuli and the level of scoring in ESP tests.

The quantitative investigations reported and referred to in the pages of this *Journal* have been mainly those in which the well-known ESP symbols have been used. Because of the fact that these have varied little in shape and size, relatively little direct insight has been achieved concerning the role of the stimulus in ESP. However, some investigators have attempted to get at this question directly by

making systematic changes in the testing materials. Carpenter and Phalen (1) found that their subjects could score as well with colors as with the ESP symbols. MacFarland and George (7) found no difference in success between the use of regular and of distorted symbols, with the notable exception of the results of one of the investigators who acted also as one of the subjects. He scored above chance on the regular symbols and below on the distorted—the effect which he had anticipated would be found. Murphy and Taves (8) used playing cards and special decks, some made up entirely of circles and blank cards, and others of circles and crosses, in addition to the usual decks of twenty-five ESP symbols, and found a tendency for the scores in various materials to vary together.

L. E. Rhine (12) varied the ESP symbols used, both as to size and as to the number of copies of each presented at each trial. In one series, she used symbols of 3½, 1⅝, ⅛, and 1/32 inches in diameter. (The measurements were all made upon the circle and the other symbols were of a proportionate size.) In another series, she compared the results from large cards stamped with a single symbol (1⅝ inches) and from cards of the same size upon which several copies of the same symbol, each 1⅝ inches in diameter, were stamped. She concludes that "within the scope of the experimentation herein reported size variations ranging to proportions of 2,704 to 1 in stimuli did not result in a significant preference. Variations in number of stimuli presented 5 to 1 at a given time did not result in a significant preference."

The present research was concerned with making further systematic tests to determine the relation of symbol size to ESP scoring. On the basis of the above-mentioned studies, there would appear to be no reason to expect differences in scoring with variations in size or shape of ESP symbols unless such differences arise from the personal preferences of the individual subject. The importance of the hypothesis suggested—that ESP is not affected by the physical characteristics of the stimulus—would require prolonged research before such a statement could be advanced as a definite conclusion. Accordingly, a more extensive investigation of ESP in relation to symbol size appeared to be fully warranted.

Two further problems arose during the course of the investigation and were considered as fully as the general plan and scope of the research permitted. One was concerned with the possible relation of the amount of experience of subjects to scoring rate. The second

involved the question of a relation between the amount of experience with a particular stimulus material and the rate of scoring. These problems seem especially apropos in view of the widespread opinion among experimenters that successful subjects decline in score averages after a period of some success. This has been noted by J. B. Rhine, (11), Pratt (9), L. E. Rhine (12), Price and Pegram (10), Gibson (3), and Riess (13), and is apparent in still earlier reports dealing with this field of research.

Restatement of the Problems

Four important problems are therefore considered in this report. They are: (1) Judging from the results of this investigation alone, does ESP occur? (2) Assuming the function of ESP, is there any relation between symbol size and the rate of scoring? (3) What is the relation between the amount of "experience" of subjects and the rate of scoring? (4) What is the relation of "newness" of material to rate of scoring?

EXPERIMENTAL CONDITIONS AND PROCEDURES

On the basis of differences in procedure, the work may be divided into two main series, both of which dealt entirely with the ESP of objects (clairvoyance). Series A was done in the period from March, 1938, to August, 1938. During this period the experimental set-up required the direct participation in the test of only one experimenter. This series was conducted by one of the writers (Woodruff) with only occasional introduction of other investigators to witness the procedure. Series B was done in the period from October, 1938, to March, 1939, and required the simultaneous participation throughout of both of the writers as experimenters. Important differences in the experimental conditions and procedures of the two series make it necessary to describe the two separately and to consider how the results of each bear upon the primary problem of the occurrence of ESP.

Series A

Subjects. Forty-two persons were tested in Series A. This number includes 14 members of Oxford (N. C.) Orphanage of high school age, 21 undergraduate students of Duke University, and 7 others ranging (in age) from adolescence to middle age.

Size of Stimulus Symbols. Throughout both series all the tests were conducted with the five ESP symbols in the usual balanced pack of 25 cards. In Series A, three sizes of symbols were used—the regular 1½ inch printed ESP symbols, ¼ inch symbols drawn in ink, and

symbols not over 1/16 of an inch, drawn in ink. The smallest size required moderately close scrutiny to decipher. The measurements in each case are approximations and are given for the diameter of the circles. The size of the cards was in all instances that of the standard playing card, with a single symbol appearing on each card face. During all tests, each run of 25 cards consisted of symbols of one size. The 1½ inch symbol was uniform in design and was printed by a commercial process in black ink. The two smaller sizes were freehand drawings made with a fountain pen in dark blue ink.

Experimental Set-up. The card matching procedure known as the STM (screened touch matching) technique was used throughout Series A. The screen, which shielded the experimenter and the deck of stimulus cards from the subject, consisted of a piece of plywood 18 inches high by 24 inches wide, held in a vertical position by means of wooden supports. It rested on the table between the subject and the experimenter, who sat opposite each other. When the experimenter and the subject were seated normally with the screen in position, each could see the top of the head of the other. Between the bottom edge of the screen and the table top was an aperture two inches high and eighteen inches long. In this opening, five ESP symbols were located in a row in such a way that they were visible to both subject and experimenter. These five key cards were chosen from the regular brown-back ESP pack which has each symbol in a different color, in order that the key cards might not be identical with any of the symbols with which the subject was to be tested. On the experimenter's side of the screen, 3½ inches back of the aperture was a low vertical screen 3 inches high and 23 inches long. Its position in relation to the aperture was such that the subject could not possibly see the cards held by the experimenter.

Procedure. With the subject and experimenter both seated and with the screen and the key cards in position, the experimenter shuffled and cut a pack of 25 cards behind the screen, out of sight of the subject. With the pack face down in his hand in readiness for dealing, the experimenter then signalled to the subject to begin by saying "all right." The subject designated his choices (as to which symbol he thought was on the top card of the deck held by the experimenter) by touching, usually with the eraser end of a pencil, one of the five key cards lying in the aperture. This response of the subject was visible through the one-way aperture to the experimenter, who immediately placed the top card of the deck opposite

the designated position but behind the second screen. Without waiting for a further signal, the subject then proceeded to touch the key card which he felt corresponded with the second card in the pack (by then the top one). The experimenter laid this card opposite the key card touched. This procedure continued until the 25 cards in the deck had been guessed.

As soon as the experimenter announced the end of the run, the subject removed the larger screen from the table. The key cards remained in position on the table. The experimenter picked up the pile of cards opposite the first key card, turned it over so that the symbols were facing upward and, while the subject watched the cards, laid those symbols which were "hits" nearer to the key cards and discarded the "misses"—at the same time counting aloud the number of hits. This procedure was followed for all five piles with a cumulative audible count of the score for the entire run. Following this, the cards segregated as hits were again examined and counted and the score was recorded by the experimenter in full view of the subject. The subject then replaced the screen and the experimenter shuffled and cut the cards preparatory to the next run. The number of runs done during a session with each subject varied somewhat with the rapidity of the subject's matching and the time at the disposal of the subject and the experimenter.

Methods of Selecting the Symbol Sizes to Be Used for Each Run. During the first part of Series A, only two symbol sizes were used—the regular 1½ inch printed symbol and the ¼ inch symbol. These sizes were alternated regularly from run to run so that the subject knew which symbol size was being used at any particular time. Likewise when the 1/16 inch symbols were introduced about midway through Series A and all three sizes were used, regular rotation among the three was followed from run to run.

However, for a short time in Series A, a variation in the alternating method of symbol selection was used. The experimenter, attempting to follow a random order in his selection, chose subjectively the cards (symbol sizes) for each run without letting the subject know until the check-up what size was being used. The experimenter restricted his choices in such a way that an equal number of runs with each size was made in each session. However, as the subject did not know exactly when the experimental period was to end, he had no dependable way of knowing which particular set of cards

would constitute the last run for the day and he was therefore limited in his ability to infer the size of symbols to be used.

During a part of Series A, the tests were conducted with a third person present to witness the entire procedure. A second member of the Parapsychology Laboratory staff was present for 169 runs, either J. G. Pratt, Miss Margaret Price, or B. M. Smith. In addition, 63 runs were casually witnessed by other subjects.

Series B

Subjects. Twenty-four undergraduate college women of Duke University and 8 other adults participated as subjects in Series B. Of these subjects, 8 had participated in Series A.

Stimulus Symbols. During Series B, four different symbol sizes were used, the three already described for Series A and a still larger size with the circle 2¼ inches in diameter. The 2¼ inch symbols were drawn by hand with a broad-pointed pen, using india ink. The characteristics of the other sizes were the same as for Series A. The size of cards in all cases was again that of a standard playing card.

Experimental Set-up and Procedure. The STM condition as used in Series B was modified in several respects intended to safeguard the procedure against possible weaknesses present in Series A.

a. The Screen. The large screen used in Series B had the same dimensions as the former one except that the aperture was 20 inches long. However, it differed from that of Series A in two respects. A small shield 5 inches in width on the experimenter's side of the screen slanted up from the table at a forty-five degree angle from a point two inches back of the opening (Fig. 1). This sloping shield permitted the experimenter who handled the cards to see the subject's choices with greater ease and at the same time it was effective to prevent cues of a visual kind reaching the subject through the aperture without requiring the use of the small secondary screen (*cf.* p. 124). The shield was attached permanently to the screen and had the additional function of serving as a rest when the screen was turned on its side on top of the table for the check-up at the end of each run. The second new feature of the screen was a horizontal row of five wooden pegs which were placed about 4 inches above the top of the aperture and on the subject's side at intervals of about 2½ inches. The key cards, which again had the colored symbols, were each punched with a small hole near each end, by means of which they were hung on the pegs. The row of pegs permitted the use of the key cards in an order unknown to the experimenter who

Fig. 1. Side view of the experimental table.

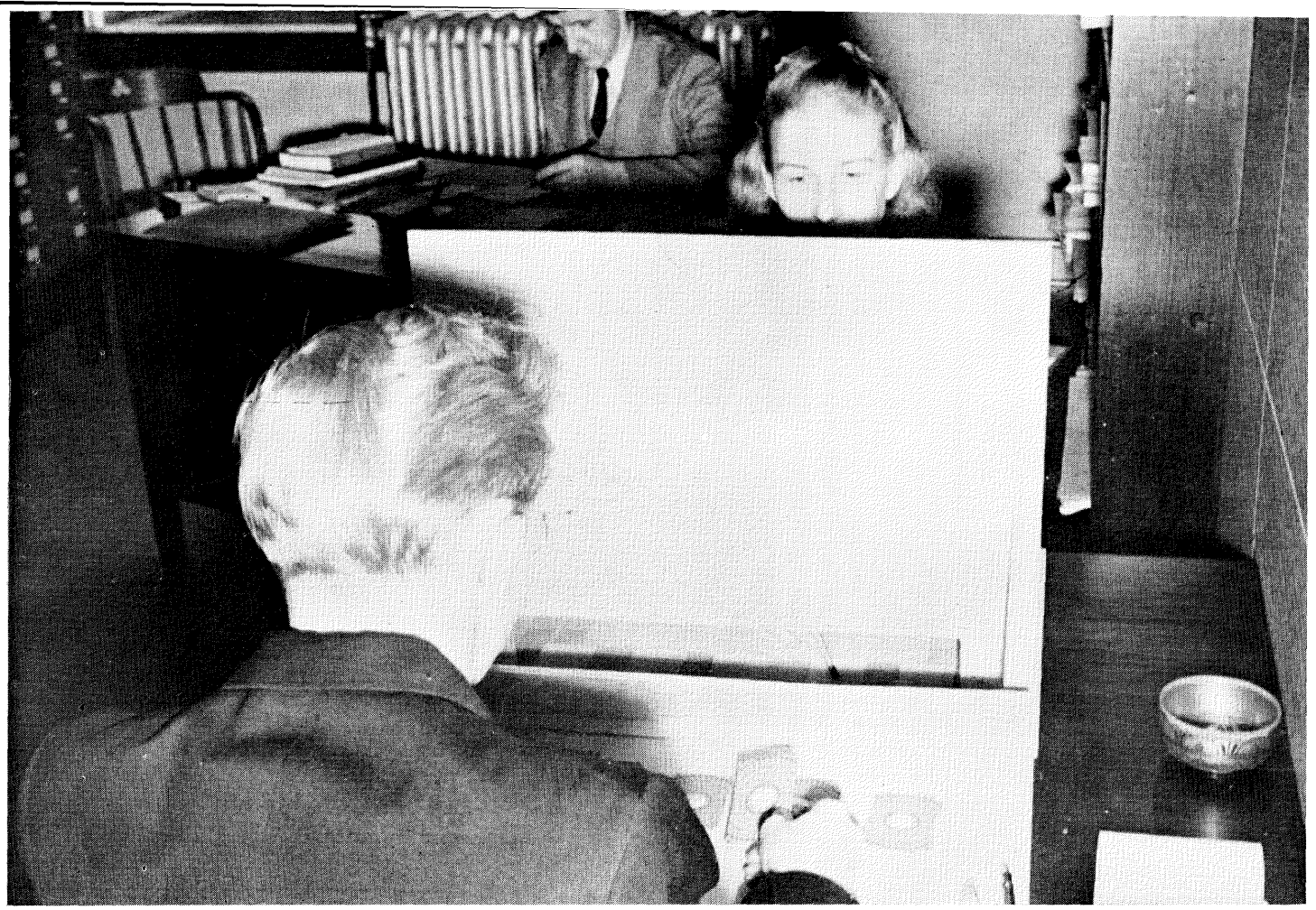

Figure 2. View of the experimental set-up from above and back of Woodruff.

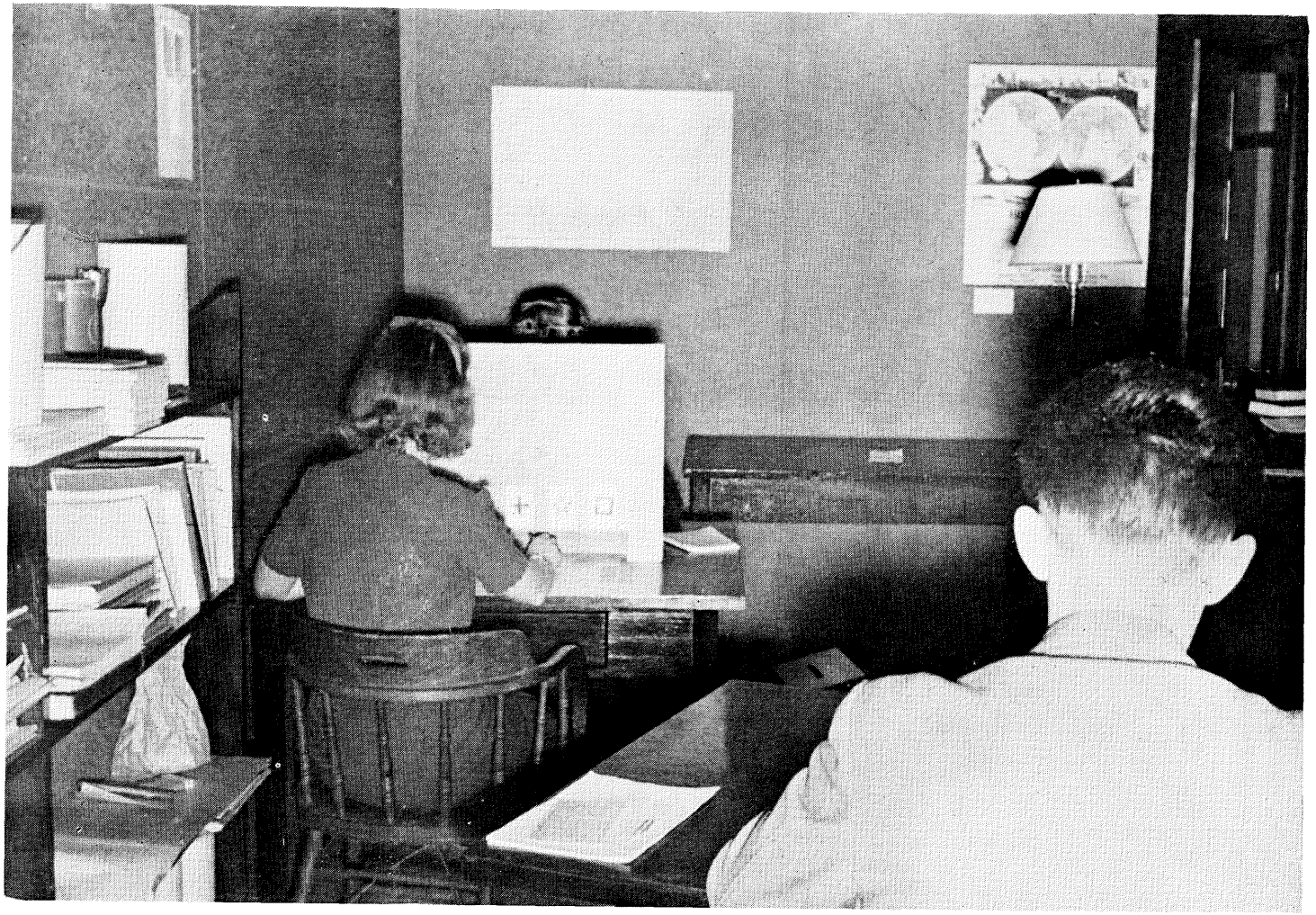

Figure 3. View of the experimental set-up from above and back of Pratt.

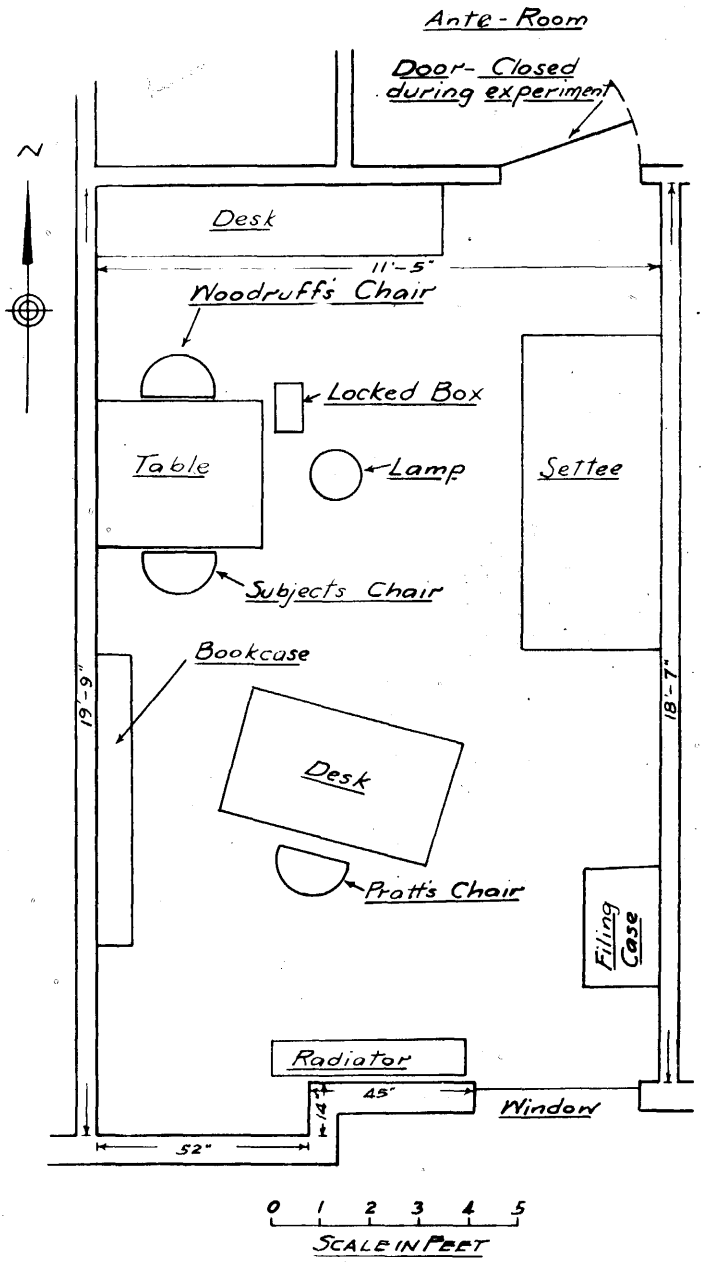

FIGURE 4. FLOOR PLAN OF THE EXPERIMENTAL ROOM

E.P.G.

handled the cards, as the key cards were arranged in a new order before each run and after the screen had been put in position. Five blank cards were placed in the aperture in position directly under the five key cards to facilitate the experimenter's dealing of the symbols in keeping with the subject's pointing.

b. Serially-numbered record sheets. Series B was broken into six sub-series involving the use of different combinations of symbol sizes, to be described shortly. The length of each sub-series was determined in advance. As a preparation for each sub-series, the experimenters wrote out a description of the length and general purpose of said series, keeping one copy for themselves and depositing another with one of the Laboratory secretaries.[1] Each experimenter was thereupon provided with exactly as many data sheets as the number of the projected runs. Each sheet had a serial number and a seal for identification purposes which was made with a special stamp available only to the secretary. The serial numbers on the record sheets of one experimenter were duplicated exactly by those of the other. The specific numbers assigned for the runs of this particular experiment were not used on any other record sheets issued in the Parapsychology Laboratory up to the time of making this report. Each experimenter was careful to use his record sheets in correct serial order. The purposes of recording required that each experimenter use one sheet for each run. Each run thus received a distinctive number at the time it was made.

c. Two experimenters: Roles during the run. The actual testing procedure for each run may be described as follows: One experimenter, Woodruff, and the subject sat facing each other across a table as in Series A. The second experimenter, Pratt, sat about six feet from and almost directly behind the subject (see Fig. 4). The screen was placed in position. While Woodruff shuffled and cut the pack of cards to be used, Pratt took the key cards from the pegs and handed them to the subject who changed their order and replaced them without giving Woodruff any indication of the new arrangement. In the last sub-series, Pratt re-arranged the key cards and put them on the pegs himself; during that period the experimenters were careful that the shuffling and cutting by Woodruff were not completed until Pratt had returned to his usual position, so that there would be no possibility of his seeing any of the cards held by

[1] The writers wish to express their thanks to Mr. E. P. Gibson for his assistance in preparing for the experiment and for his independent re-checking of all the results.

Woodruff after they were shuffled. Woodruff then gave the signal to the subject to start, and the subject proceeded to indicate his "guesses" by pointing to the blank cards in the opening under the screen. Woodruff distributed the cards, following the subject's pointer, but he was in complete ignorance throughout the run of the symbol designation intended by the subject.

d. Recording. At the end of the run, the screen was left in position on the table while Woodruff recorded the actual distribution of the 25 cards on the appropriate record sheets, and while Pratt recorded the order of the key cards on his record sheet bearing the same number. The order of key cards was recorded by Pratt in reverse order so as to make them correspond with Woodruff's record when the two sheets were juxtaposed later for checking. Pratt in addition recorded the name of the subject, the type of test, the date, and the initials of the experimenters. This recording was done without any communication between the experimenters or from the subject.

e. Locked record box. When Pratt finished his record, he carried it to the experimental table. Woodruff had usually finished his recording by this time. In case he had not, Pratt was careful to keep his record out of Woodruff's visual field until the other record was completed. Woodruff then clipped together the two independent records with the common serial number and deposited them without further marks or observation of the sheets themselves in a special locked box provided by the secretary for the purpose.

f. Counting. The screen with the key cards still on the pegs was then laid on its side, either by the subject or by Pratt, so that both the key cards and the 25 cards as distributed were visible to all three persons. Pratt then proceded to sort out the hits from each pile, laying them nearer the key cards and counting aloud the number of hits for the run. This process was observed by Woodruff and the subject. The hits as segregated were then re-examined and re-counted. The score for the run as thus determined at the time from the cards themselves was recorded immediately by each experimenter in his personal record book.

To continue the test, the screen was again raised to its vertical position, the key cards were re-arranged upon the pegs, Pratt returned to his seat behind the desk, and Woodruff, having shuffled and cut the pack of cards, gave the subject the signal to begin.

g. Experimental periods and rates of performance. As in Series A, the length of each experimental session was not fixed but was adjusted

to suit the convenience of the subjects. The speed of work varied somewhat from subject to subject, but for the average the number of runs performed within an hour was about thirty. Usually the subjects worked by appointment for from thirty to forty-five minutes at a session.

h. Obtaining subjects: Degree of selection. The first subjects to be used in Series B were those of Series A with whom the experimenter was able to make contact and who were interested in continuing the investigation. From time to time these subjects suggested the names of interested friends and in this way a considerable number of new subjects were brought in. No particular effort was made to select subjects on the basis of their performance or excellence in the tests. In general, however, those who did better were more interested to continue and were encouraged to do so. As far as possible, each subject worked one period each week.

i. Checking the scores from the record sheets. The record sheets were checked entirely independently by the secretarial assistant. After he had obtained the scores by juxtaposing the key cards (as recorded by Pratt) with Woodruff's record of the symbol distribution, a run by run comparison of the scores as recorded by the two experimenters at the end of the run and as found from this re-check of the record sheets was made. In case of a discrepancy, the written records were consulted again immediately to see whether the difference could be accounted for. If the difference were evidently an error from checking the record sheets, the secretary's score was adjusted to correspond to that of the experimenters. If, on the other hand, the re-check of the record sheets did not account for the difference, the lower score was accepted as the official one and was entered thereafter in all computations from the data.

Selection of Symbol Size to Be Used. During the first sub-series of 300 runs, the 1/16 inch, the 1/4 inch, and the 1½ inch symbol sizes were selected subjectively by Woodruff in an attempt to approximate a random order, as for part of Series A. In the next 600 runs, the same symbol sizes were used, but the choice of the size for each run was determined by the cast of a die. Until this point in Series B, the subjects did not know until the end of the run what size of symbol was being used at any time.

In the next 300 runs only the regular 1½ inch symbols were used. During the last 1,200 runs, the 1½ inch and the 2¼ inch symbols were used alternately. During this period Woodruff was careful to

inform the subjects which size was being used before they began their responses for each run.

Special Points of Procedure: a. Knife-cutting. During the last 830 runs of the experiment, the cards were shuffled by Pratt and were cut by Woodruff by means of a paper knife. The object of this variation in procedure was to determine whether extra-chance scoring might depend upon inadequate shuffling or upon peculiarities in the cards which make them cut by hand in a non-chance manner.

b. Shuffling methods. Woodruff used the method of shuffling in which the pack is held in one hand while cards are slipped out of it and re-inserted into the pack with the other. Pratt, on the other hand, divided the pack somewhere near the center and riffled the two halves together in a manner which, superficially, would appear to give a more adequate mixing of the cards. This was repeated five times for each deck of cards. Cutting with a paper knife theoretically permits of no effect of warping in favoring a division of the deck at particular points more than others.

c. Inverted keys: Blind STM. In the last sub-series of 400 runs, the key cards were placed facing inward upon the pegs so that the backs only were visible to the subject. During this series, the rearrangement of the key cards was always done by Pratt. The subject did not know the order of key cards, unless some of the symbols were recognized by cues from the backs of the cards. The purposes of the experimenters in making this innovation were, first, to introduce a novelty into the situation which might add to the interest of some of the subjects, and second, to provide an easy step toward a new experiment beyond the ones here reported which, it was feared, the subjects might consider to be too difficult without some transition.

Summary of Procedure: Series B. To help the reader fix in mind the experimental procedure for Series B, the essential steps of the plan of investigation may be reviewed: (1) Both experimenters were present during all the tests, each with a definite, pre-assigned role to facilitate the procedure and to safeguard against experimental error. (2) The subjects were tested for their ability to guess cards completely screened from sight and handled entirely by Woodruff. (3) The subject indicated his guesses by pointing in relation to five key symbols which were out of Woodruff's sight and unknown to him until after he had recorded the 25 cards as distributed at the end of the run. (4) Meanwhile, Pratt, without seeing Woodruff's cards, recorded the key cards and other essential data about the run.

(5) The two separate records were deposited at once in a locked box to be scored later by a third person. The record sheets were serially numbered in pairs and designated for the purpose of the investigation so that every run had to be clearly accounted for. (6) The two experimenters jointly checked each score from the cards and each entered the number of hits observed in his personal record of the run scores as obtained at the end of each run. (7) The laboratory secretary independently checked the scores from the record sheets. His scores were compared with those of the experimenter and in case of a discrepancy not immediately accounted for, the lower score was adopted.

EVALUATIVE PROCEDURES

The statistical methods used in the evaluation of the results are, in the main, standard procedures. The critical ratio method, the chi-square method, and the method for the evaluation of a difference are conveniently described with specific reference to the data of ESP research by Greenwood and Stuart (5). On the strength of Greenwood's empirical findings (4), the results were evaluated on the binomial hypothesis.

Methods of correcting a P-value derived by the critical ratio method for the possible factor of optional stopping—i.e., taking advantage of the trend of scoring throughout the experimental series to stop the tests when the total results "favor" a particular interpretation—have been devised by both T. N. E. Greville and J. A. Greenwood. A description of the latter's method, which has been applied in the present study, is awaiting publication, and a full explanation cannot be undertaken here. In general terms, the applicability of the method is based upon the assumptions that the total experimental series to be evaluated consists of sub-series having two characteristics: (1) a stated maximum number of such sub-series beyond which the experiment would not go; and (2) a fixed length for each sub-series. It is important that neither of these characteristics be influenced by the results, a requirement which is most clearly met if they are determined before the experiment is started. This is not to say, however, that otherwise the essential conditions for applying the optional stopping method are necessarily lacking. If it can be established that the maximum number of sub-series and the length of each one were not affected by the preceding scores in the experiment, that is all that is required to make the optional stopping method applicable. When the probability of chance occurrence for the results

from the beginning of the experiment to the end of a particular sub-series is obtained, the optional stopping correction converts this value into the probability of a chance occurrence of the same deviation ratio at the end of any one of the possible stopping points (end of each sub-series). The application of this method to the results of Series B is discussed later.

DEFINITIONS

For purposes of the presentation of the results and discussion in later sections of the paper, the following definitions are given.

Those subjects will be designated as "experienced" subjects for a given series (A or B) who had participated previously in formal ESP tests, irrespective of the size of symbols used, either with the writers or with any other ESP investigator. All other subjects in each series will be "inexperienced." Thus an "inexperienced" subject in Series A was, if he continued through Series B, "experienced" in the latter. A particular subject is considered as maintaining throughout a given series the status of "inexperienced" or "experienced" with which he began that series.

"New" materials for a given subject in a given series are those symbol sizes with which he had not been tested previous to the series in question. Other material will be designated as "old" material. Thus "new" material in Series A becomes "old" material when used by the same subject in Series B.

RESULTS

I. AS EVIDENCE OF ESP

The Evaluation of the Results in Relation to the Hypothesis of Chance Coincidences

The Experiment as a Whole. This report deals with a total of 3,868 runs with ESP cards, or 96,700 trials, each trial with a probability of success of 1/5. The successes, or hits, observed were 20,310. This number represents a deviation from mean chance expectation of 970 hits, or an average of 5.25 hits per run. The standard deviation (S. D.) of expected hits for this number of trials is 124.38, and the critical ratio (C. R.) of the result is 7.80.

Series A. Forty-two subjects participated in Series A with a total of 1,468 runs in which they scored 7,821 hits. This is 481 hits in excess of mean chance expectation or an average number per run of 5.33. The S. D. for this number of runs is 76.63. This gives a C. R. of 6.28 with an equivalent probability value (P) of 10^{-10}. The total of 169 runs witnessed by another staff member in addition to

Woodruff gave a positive deviation of 73, with an average of 5.43 and a C. R. of 2.81.

TABLE I

SUMMARY AND COMPARISON OF GENERAL EXPERIMENTAL RESULTS OF SERIES A AND SERIES B AND THE RESULTS OF THE CROSS-CHECK ON SERIES B

Series	Runs and Dev.	Av. Hits per Run	S.D.	C.R.
A	1,468 + 481	5.33	76.63	6.28
B	2,400 + 489	5.20	97.98	4.99
Total	3,868 + 970	5.25	124.38	7.80
	C.R. of diff. = 2.00			
Witnessed Tests in Series A	169 + 73	5.43	26.00	2.81
Cross-Check Series B	2,400 + 56	5.02	97.98	.57

Series B. Series B consisted of a total of 2,400 runs, or 60,000 trials, of which 12,489 were hits. This is a deviation of 489 in excess of mean chance expectation, or an average of 5.20 hits per run. The S. D. is 97.98, which gives a C. R. of 4.99 with the associated probability of 3×10^{-7}. An analysis was made of the 2,400 runs of this series by the chi-square method. This analysis, based upon the frequency of run scores for all the subjects as shown in Table II, gave a chi-square of 34.30, with 9 degrees of freedom and a probability of .000,078. Thus the deviation ratio method of evaluation and the method of chi-square both support the conclusions that results reliably different from chance expectation were obtained in Series B.

It is evident from the general summary in Table I that the results of the research as a whole and of the two series taken individually can not reasonably be attributed to chance factors. Because of the more elaborate precautions against sensory cues and experimental error which were taken in Series B, more interest attaches to the results of this series as regards the question of the interpretation of the deviations. It seems important to reach some kind of conclusion as to whether these results were due to ESP before proceeding to a consideration of further problems dealing with the nature of ESP.

The Results in Relation to the Problem of the Occurrence of ESP

Series B. Series B was planned shortly after the symposium on experimental methods in ESP research at the Columbus, Ohio, meeting of the A.P.A., in September, 1938. In planning their research, the investigators made every effort to take fully into account all the criticisms of methods made at the symposium, as well as those in

TABLE II

FREQUENCY DISTRIBUTIONS OF RUN SCORES FOR SERIES B BY INDIVIDUAL SUBJECTS AND WITH THE TOTAL DISTRIBUTIONS FOR THE EXPERIMENTAL SERIES AND THE CROSS-CHECK

	Frequency of Run Scores															*Runs and Dev.*
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Subjects																
H.G.	1	1	8	18	37	36	34	27	13	7	4	0	1	0	0	187: + 76
M.B.	1	1	11	26	27	30	28	14	13	3	2	2	0	0	0	158: + 8
B.Y.	1	1	6	12	18	14	7	5	8	4	2	0	0	0	0	78: — 2
J.Bd.	0	1	3	7	13	13	9	9	2	1	1	0	0	0	0	59: + 2
D.L.	0	1	11	10	17	29	19	14	8	4	4	1	1	0	0	119: + 46
O.M.	0	0	1	2	7	3	0	0	0	2	0	0	0	0	0	15: — 6
A.B.	0	7	6	13	23	21	17	17	4	2	0	0	1	0	0	111: — 17
A.M.	2	9	18	45	47	68	41	34	28	14	4	0	3	0	0	313: + 53
B.M.	1	2	8	11	22	20	17	11	12	3	1	1	0	0	0	109: + 17
L.D.	0	1	7	11	19	12	10	13	3	5	0	0	1	0	0	82: + 6
J.Br.	0	0	0	3	0	0	1	2	0	0	0	0	0	0	0	6: — 1
M.E.	0	1	0	0	5	3	1	3	2	0	0	0	0	0	0	15: + 4
R.K.	0	2	0	6	6	6	5	4	3	2	0	0	0	0	0	34: + 4
C.W.	2	6	5	9	13	14	14	10	3	2	1	0	0	0	0	79: — 24
E.G.	0	1	1	2	0	1	3	1	1	0	0	0	0	0	0	10: — 3
P.M.	0	2	7	18	25	26	25	24	9	10	11	4	1	0	0	162: + 136
B.Br.	0	1	2	9	10	7	8	9	3	0	0	0	0	0	0	49: — 3
C.H.	0	0	2	0	4	2	2	4	1	2	0	0	0	0	0	17: + 11
J.A.	0	2	0	5	3	10	3	2	3	2	0	0	0	0	0	30: + 3
B.J.	0	0	2	2	5	8	5	3	1	2	0	0	0	0	0	28: + 7
D.S.	0	0	1	1	4	0	3	3	1	0	0	0	0	0	0	13: + 3
J.B.	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	2: — 3
G.E.	0	1	2	4	6	7	7	3	0	0	0	0	0	0	0	30: — 11
B.B.	1	4	7	17	32	33	24	22	9	4	1	1	0	0	1	156: + 23
C.C.	0	3	10	23	34	38	33	25	11	8	5	4	0	0	0	194: + 75
T.E.	0	0	0	1	0	4	3	1	0	1	0	0	0	0	0	10: + 7
D.A.	0	0	2	6	2	12	10	9	3	4	2	0	0	0	0	50: + 43
M.W.	1	0	2	9	6	10	13	7	1	0	0	1	0	0	0	50: + 1
N.A.	0	0	8	17	24	24	17	20	10	3	2	1	0	0	0	126: + 33
N.S.	1	2	6	8	9	12	8	5	9	5	0	1	0	0	0	66: + 15
D.C.	0	0	0	4	4	6	6	0	1	1	0	0	0	0	0	22: + 1
C.K.	0	2	2	1	9	2	1	1	1	1	0	0	0	0	0	20: — 15
Total	11	51	138	301	432	471	374	302	163	92	40	16	8	0	1	2,400: + 489
																$X^2 = 34.30$; P = .000,078
Cross-check frequency	4	64	165	323	454	464	379	268	158	80	33	4	3	1	0	
																$X^2 = 6.20$; P = .86

the literature of the critical writers. In addition, efforts were made to anticipate criticisms which had never been made and which might never be seriously advocated.

In general, the criticisms directed against the published ESP reports have been classified as those pointing to the possibility of sensory cues in the experiment; those concerned with the occurrence

of experimental errors in the observation of responses, recording results, and reporting the data; and finally those dealing with the methods of evaluation. The conditions which obtained in Series B may be considered in relation to these three general aspects of the research.

TABLE II B

CHI-SQUARE EVALUATIONS FOR INDIVIDUAL SUBJECTS (AS SUGGESTED BY THE REVIEW COMMITTEE) AND INDIVIDUAL CRITICAL RATIO EVALUATIONS

Subject	X^2	d.f.	P	C.R.	P
H.G.	9.51	8	.30	2.79	.0026
M.B.	3.49	7	.83	.32	.38
B.Y.	8.81	7	.27	— .11	.46
J.Ba.	2.62	4	.63	.13	.45
D.L.	8.61	7	.28	2.11	.018
A.B.	12.65	7	.08	— .81	.21
A.M.	13.13	9	.16	1.50	.067
B.M.	5.35	7	.62	.81	.21
L.D.	4.46	6	.62	.33	.37
C.W.	17.51	7	.015	—1.35	.089
P.M.	75.63	9	.000000	5.34	4.6×10^{-8}
B.Br.	0.85	4	.92	— .21	.42
B.Bc.	4.34	8	.82	.92	.18
C.C.	13.48	8	.10	2.69	.0036
D.A.	7.43	4	.12	3.04	.0012
M.W.	4.60	4	.33	.07	.47
N.A.	5.04	7	.66	1.47	.071
N.S.	12.01	7	.10	.92	.18
Misc.	5.77	8	.67	.00	.50
Total	215.29	128	.00000083*		

*Derived by the use of the formula:
C.R. $= \sqrt{2X^2} - \sqrt{2(\text{d.f.} - 1)}$

The Question of Sensory Cues. In Series B there was no direct sensory contact between the subject and the cards to be guessed. This aim was simply and effectively accomplished by removing the cards from the hands of the subject and by interposing the opaque screen between the subject's eyes and the shuffled pack. If sensory cues affected the subject's scores, therefore, this would have had to come about by some indirect or more subtle means. The possibilities may be examined in the light of the actual conditions.

a. Visual. Visual perception must obviously be controlled if the conditions are to be adequate for testing ESP, since the stimuli and cards are characteristically visual objects. One thinks first of possible reflecting surfaces in which the subject might have seen the card symbol. The top of the table on which the subject worked was

covered by a blotter which would have prevented reflection even without the screen; and with the small shield back of the aperture of the screen, there could be no possibility of reflected visual cues. The walls of the room in which the experiment was conducted were of soft composition material and were equally poor as reflectors. The subject could not have made a practice of looking over the screen into Woodruff's eyes or glasses without having had his actions detected by one or both of the experimenters, one of whom sat behind the subject with the latter in full perspective.

One critic has suggested that in the usual form of the STM procedure in which the key cards are visible to both the subject and the experimenter, the experimenter may wishfully misplace a few cards in such a manner as to get the scores which he anticipates. This assumes that the experimenter may use sensory cues to produce spurious results. The conditions of Series B explicitly prevented this danger by having the subject alone know the order of the key cards. Any prejudice or will-to-produce on the part of the experimenter who handled the cards was effectively controlled.

b. Auditory and visual. If cues occurred, therefore, they must have been partly of an auditory character, effected with the following assumed steps: The experimenter either deliberately looked at the cards in the deck or unwittingly observed cues that identified them. He unconsciously or deliberately gave cues to the subject that could be heard by him but not by the other experimenter seated a few feet behind him. The subject would follow the cue and point to the key card indicated. Or the subject might give the experimenter auditory cues concerning the order of the key cards which would permit the experimenter to misplace some cards to increase the score. This is to assume, again, that the experimenter either looked at the faces of the cards or identified them through visual or tactual cues.

The facts as they bear upon these possibilities are these: (a) The experimenter gave no signal to the subject throughout the run other than that of the time to start. This is one of several points in which the methods of this study exceed the requirements for control against sensory cues and experimental error as laid down by Knight Dunlap and others (2) in their description for the condtions of an adequate experimental testing of the ESP hypothesis by means of card sorting. They suggested that the experimenter should give some vocal signal when he was ready for each trial throughout the run. (b) The rate at which the subjects proceeded in their indi-

cations in the present research averaged for some of the best scorers as fast as two cards per second. This in itself would appear to be an effective block against the interchange of auditory signals between experimenter and subject. (c) As later analyses of the data will show, subjects tended to decline in their ability to score above chance the longer they used a particular kind of stimulus material. This is a fact which is difficult to account for on the basis of the use of sensory cues, either visual or auditory. On the other hand, a decline in ability to demonstrate ESP has frequently been reported, even by investigators in experiments done at such a distance that the question of auditory and visual cues could not enter.

In certain respects, however, our conditions failed to meet the requirements laid down in the paper mentioned: (1) One specification was that the scores be withheld from the subject throughout the entire research. In this experiment, the subject knew his score at the end of each run of 25 trials. (2) The order of the key cards was to remain unchanged throughout. In our tests the order of the key cards was changed from run to run. (3) No computation of scores was to be made until the end of the experiment. They were frequently made in this experiment. (4) Each subject was required, as far as possible, to have the same number of tests as every other subject. No effort was made in the present work to obtain the same number from each subject. (5) Work periods were to be of uniform time length, were to consist of the same number of runs, and were to have the same distribution throughout the week. While the work periods in this experiment were roughly uniform as a matter of convenience, there was no effort made to keep the other points uniform. (6) Age range was to be restricted, for example, to two adjacent college years. Our subjects varied more widely than this and no effort was made to restrict age. (7) Subjects were to be requested not to use ESP cards in any other connection during the course of the experiment. No such request was made in this research. (8) A number of statistical requirements were made that were not carried out in this particular study. For example, the scores of each subject were to be totalled (a) for each set of five successive runs, (b) for each successive set of 25 runs, (c) for all the tests of each work period, (d) for the total tests of the experiment. This was done only for (c) and (d). The requirement stated that there should be tabulation of the total hits and misses for each of the stacks separately; that is, for each key card. This was not done. There

was required also the average number of successes per subject, average score for experimental session, and percentage of successes. (9) It was required that the experiment be set up under the superintendence of three psychologists, each from a different university. It was, in addition, to be under the direction and control of two or more psychologists who are regarded by members of the profession generally as competent in the experimental field, one of whom was to be on duty during every work period. In this experiment there was no superintendence from psychologists of other universities, and since in the less objective professions, competence of the experimenter is mostly determined by the *a priori* acceptability of his findings, it is conceded that in this point, too, the requirement is not met.

While the majority of the requirements which have not been met can be recognized to have a certain value for experimental objectives not concerned here—objectives such as the comparative study of ESP test performance under certain conditions—we are unable to discover any reason warranting their general adoption. The proponents themselves gave no grounds for their being regarded as essential to a crucial test of the ESP hypothesis. The ninth requirement, regarding superintendence, is based upon the assumption that "competent" experimenters will remain "competent" should they become associated with an investigation in which the findings are favorable to the ESP hypothesis. So far as is known, there was never any question of the competence of the now considerable number of psychologists who have obtained results favorable to the hypothesis prior to their publication of these findings.

(d) The series of 400 trials made with the key cards facing inward, so that only their backs were visible to the subject, constitutes in a peculiar way a control upon the possibilities that sensory cues of visual and auditory characters may have been combined in producing the results. While it can not be denied that the subject may have identified some of the key cards from their backs, there will be no question that the subjects knew far less about the order of key cards in this series than in the others in which the symbols were fully visible. Consequently, if the subject was using information obtained visually concerning the key cards and was conveying such information to the experimenter as would permit the experimenter to misplace certain cards as visually perceived through cues on their backs, the results of this series should have been appreciably lower than those of the remaining 2,000 runs. Actually, the scores with blind STM

were at the same average level as for the rest of Series B, an average of 5.2 per 25.

c. Faulty cards and shuffling defects. Any effect of inadequate shuffling or faulty card materials which permitted the subject to use inference to score above chance would indirectly involve sensory perception. This is true in the sense that the subject would have to apply knowledge which had previously obtained through the senses to infer something of the actual order of cards after they had been shuffled. This possibility was explicitly controlled during more than a third of Series B, during which Pratt shuffled the cards and Woodruff cut them out of sight behind the screen with a paper knife. The 830 runs done under this condition gave a deviation beyond mean chance expectation of 177 hits, an average of 5.21, almost exactly the same as that for the series as a whole.

Safeguards Against General Experimental Errors: a. Recording and checking. The possibility of errors in recording the results was avoided by the simple expediency of having Woodruff record the cards as distributed, and Pratt the key cards, without either one having any knowledge of the observations of the other till his record was fully made. The third person who checked the record sheets later did so without any knowledge of the scores as obtained from the cards after the run and recorded by each of the experimenters. Some light is thrown on the question of the accuracy of the two methods of checking, either from the cards or from written records of the symbols, by the following facts: When the scores as obtained from the record sheets were compared with those of the two experimenters as obtained from the actual cards, several discrepancies were found. In most instances in which the record sheets were again consulted, it was immediately evident that the error in checking had been made by the person working from the written records. In three instances it was evident from the record sheets that an error in recording had been made. One of these consisted in recording two key symbols of the same kind in the series of five where all were known always to be different. The other two were evident from a study of the symbol frequencies in the card distributions which showed that six of one symbol and four of another had been recorded for one run, when only balanced packs of five symbols of each kind were used. All told, three *recording* errors were discovered; that is, errors in which one of the experimenters had made a mistake in writing down the symbols. No errors in counting the scores from

the cards at the end of the run were detected. Two of the recording errors lowered the run score by one hit in each instance, and the other raised it by one hit. The net result upon the total deviation as represented by the experimenters' scores was, therefore, to lower it by one hit for the entire 2,400 runs.

b. Deception. Experimental conditions which would make it impossible for one investigator wilfully to deceive his colleagues might not be attainable. However, it is worth pointing out that the conditions of Series B accomplished something in this direction, inasmuch as they made it difficult, if not out of the question, for one experimenter to practice deception upon the other even if he had wished to do so. The serially numbered record sheets which were obtained from the secretary for the purposes of the experiment were stamped with a seal which was always locked in the secretary's keeping and which could not have been duplicated or "borrowed" by either experimenter without considerable difficulty. The presence of the locked box into which the record sheets were deposited at the end of each run would have made it necessary, even if an experimenter had succeeded in duplicating the blank record sheets, for him to recover the legitimate record before substituting a faked one, or to recover the legitimate record before the check-up by the secretary if he intended to change it in a way to improve the scores. Finally, each experimenter kept his own complete record of the run scores as counted. If either one had wished to change the results, it would have been necessary for him to secure the record of the other and change this as well. These, it would appear, may not be insurmountable difficulties, but they are real psychological barriers to dishonest practices which those who wish to consider the question of fraud on the part of the experimenters may want to take into account.

The Question of Proper Statistical Evaluation

a. Sampling. The point most generally raised here is that of whether the data as evaluated were properly selected. In particular, were all the scores observed included in the final evaluation of the experiment? A positive answer to this question for Series B is particularly easy and emphatic because of the use of serially numbered record sheets. As stated previously, the length of each sub-series was definitely fixed in advance; the general procedure to be followed was outlined and the descriptive statement given to the secretary at the

time the required number of serially numbered sheets was requested. In this way the investigators put themselves on record at each new stage of the research as to the additional number of runs that would be made. Each run was recorded on the sheet provided for the purpose before anything was known as to the actual score. In the final check-up from the written record, all of the blank sheets, duly filled in and signed, were accounted for. Any omissions would have been immediately obvious. There can be little question, therefore, that the 2,400 runs reported represent a consecutive series which was all the work done under the conditions described by the two experimenters within the time limits stated.

b. Optional stopping. Actually, the point at which the present experiment was arbitrarily terminated for the purposes of making this report did not represent an end of joint investigation by the two writers. The stopping point was determined actually by the occasion of presenting a report to the Southern Society for Philosophy and Psychology. The further work was subjected to alterations of conditions not relevant here. This raises a question as to whether the experimenters simply selected a favorable point at which to close the investigation, a point for which the only statistically reliable factor was their exercise of that right of optional stopping. The effect of optional stopping as related to ESP data has been emphasized by Leuba (6). The mathematical aspects of the problem have received particular attention from Greenwood, whose method was described in general terms earlier in this report. The optional stopping correction was discovered as the present research was nearing completion. The length of each sub-series as well as the maximum number of sub-series which would be done had not been stated before starting Series B. Before applying the optional stopping correction, therefore, it was necessary to make sure whether these characteristics were affected by the experimenter's knowledge of the scores throughout the experiment.

The lengths of the six sub-series in Series B were 300 runs for the first four and 1,200 and 400 runs for the last two, respectively. Actually, (a) the average remained fairly constant throughout the experiment; and (b) the experimenters were not aware of the manner in which either shortening or lengthening the sub-series would bear upon the evaluation of the results. These facts satisfy the writers that they were not influenced by the scores in fixing the length of sub-series. However, it was agreed that for the purposes of correcting

for a possible effect of optional stopping, Series B would be treated as though it had consisted of 8 sub-series, each of 300 runs. In this manner the possibility of favorable variations in length are completely ruled out.

Likewise, no maximum number of sub-series for the experiment as a whole had been set at the start. Under the circumstances, the investigators, in order to make a fair correction for the effect of optional stopping, arbitrarily set the outside limit, beyond which the experiment would not have proceeded under any circumstances, as twenty sub-series. The two experimenters would need to work together intensively for one and one-half additional years in order to reach this maximum. This limit was deliberately placed high in order that the fullest allowance might be made for the effect of optional stopping. When Greenwood's correction is applied to the results of Series B, the probability is increased from 3×10^{-7} to 5×10^{-6}.

c. Cross-checks. In order to see what results would be yielded by the actual card distributions and by the key card orders of Series B if extra-sensory perception were ruled out, a cross-check between the card distribution of a particular run and the key card order for the third run in advance was made. The data were cross-checked within groups of 100 runs, following the system of checking the first distribution of card symbols against the order of key cards for the fourth run, the second distribution against the fifth order of key cards, etc., and finishing up each group by checking the ninety-eighth distribution against the order of key cards for the first run, the ninety-ninth against the second, and the one hundredth against the third. These 2,400 empirical scores gave 12,056 correspondences, or a positive deviation of 56 from mean chance expectation and an average of 5.02 hits per run.

A chi-square evaluation of the frequency distribution of scores obtained on the cross-check shows no significant departures from chance expectation. The chi-square was 6.20, with 11 degrees of freedom, with equal probability that 86 in 100 such empirical series would on the average give a worse fit to the theoretical binomial curve.

The results of the deviation ratio evaluation of the cross-check scores are shown in Table I, and those of the chi-square treatment of the frequency distribution of cross-check scores in Table II.

d. Care in statistical treatment of the data. Related both to the

question of the general trustworthiness of the investigators and to that of the accuracy of statistical evaluations is the amount of care used in compiling and making computations from the data. The general summary of the results of Series B were calculated independently by two persons. Likewise, the cross-check with the written records of Series B was made independently by two persons, with a run-by-run comparison of their sets of 2,400 empirical scores and a final joint examination of the original records in cases of disagreement. The actual computations for Series A were the responsibility of one of the experimenters (Woodruff); but in making the further analyses of the data soon to be presented there was ample opportunity to check upon the accuracy of the figures. That is to say, the consideration of the data along the lines of various testing conditions offered a check both upon the general totals and upon the accuracy of the analyses themselves, inasmuch as the records were always retotaled and compared, after making any particular study of the results, with the general totals from which the divisions started.

e. "Stacking error." A conceivable source or error which will be called the "stacking error" may be described as follows: Woodruff, in laying off the cards of the pack, may have used either sensory cues or wilful deception to group the cards, laying an unusual number of like symbols in each pile. This is to suppose, of course, that he accomplished this by failing to follow exactly the subject's pointer. On the assumption that the experimenter actually did group the symbol suits as he laid down the cards, the element of chance in that step of the test is removed. Each run then reduces, in effect, to *five* trials in which the piles are compared with the key cards. Because of the fact that the experimenter is ignorant of the actual order of key cards, such a grouping of the symbols would not affect the average score expected. It would, however, increase the variability of the run scores, so that the probability of both high and low scores would be increased. The question at issue is whether the average of 5.2 per run for Series B would be significant in the face of the hypothesis that just such illegitimate groupings of the card symbols occurred with an indeterminate, equivalent reduction in "trials." (If this interpretation were preferred and the result were nevertheless shown to be significant, we would have to suppose that Woodruff had demonstrated ESP in favorably locating the piles.)

The results of the chi-square evaluation of the score frequencies obtained in the cross-check permit the definite conclusion that no

unusual grouping of the symbols as they were distributed occurred. For if this had been the case, these arrangements would have affected the scores of the cross-check in the very manner which the hypothesis in question supposes was the case for the experimental series. The absence of symbol groupings is demonstrated by the fact that the chi-square evaluation for the cross-check gave a P of .80 (see Table II).

It is recognized that this statistical control is one which might break down in tests in which much higher averages, necessitating some degree of grouping to produce the observed scores, are obtained. However, in the present investigation it was effective because of the fact that the average rate of scoring over a relatively long series was not high enough to be associated with noticeable symbol groupings.

From the consideration of the results of Series B in the light of all the experimental conditions, the writers are unable to offer any explanation of the findings except to say the subjects demonstrated a degree of success in identification of the concealed cards and that this knowledge was not obtained through any recognized sensory channel. On the basis of the joint investigation in particular, a conclusion is reached that ESP occurred in this investigation.

The Occurrence of ESP—Series A

The question arises, of course, as to whether the introduction of the advances in methodology in Series B means that the writers intend to minimize the importance of the results of Series A or of earlier experiments in general. The answer must be that the following considerations appear to make such a course unnecessary.

In the first place, Series B effectively substantiates Series A. In each case, the results are shown to be highly significant—Series A giving a C.R. of 6.28; and Series B a C.R. of 4.99. The average of Series A is .13 higher than that of Series B, but, as shown in Table I, this is not a reliable difference (C.R. of the diff. = 2.00). Having been forced, for want of any other explanation, to conclude that ESP occurred in Series B, the writers prefer on the principle of parsimony of hypotheses to extend this conclusion to cover Series A as well. Actually, the only differences in conditions between the two parts of the investigation are the absence of a second observer throughout Series A and a reliance therein upon counting and re-counting the scores from the cards without making independent written records. However, 169 runs of Series A with an average of 5.43 were witnessed by a second experimenter. The absence of written records in Series A does not seem so serious in the light of the comparative study of the

relative efficiency of methods of scoring in Series B which showed counting from the cards to be more accurate than the use of written records.

As long as the question of the occurrence of ESP was primarily at issue, Series B was rightly considered to represent a higher plane of evidence because of the more advanced experimental conditions. Now that the evidence on that first problem seems to justify further study of the results to see how varying the conditions of the experiment affected this phenomenon, the results of the entire investigation are admissible as evidence bearing upon possible relations of ESP.

II. RELATIONS SHOWN BETWEEN CONDITIONS AND RESULTS

The Relation Between Rate of Scoring and Size of Symbols

It was stated above that the primary purpose with which Woodruff undertook the tests described as Series A was that of making further observations upon the comparison of level of scoring and the size of symbols. The results bear out the earlier finding of L. E. Rhine (12); namely, that no significant relation is indicated.

The analysis of the *total* results of the entire research giving comparison of the scores for the four sizes of symbols used (Table III) shows no significant difference in the level of scoring for different sizes of symbols. It may be seen from the table that marked differences in averages resulted *in Series A*; total of 576 runs with the regular-sized 1½ inch symbols averaged 5.15 hits per run, while 691

TABLE III

RESULTS OF SERIES A AND B ACCORDING TO SYMBOL SIZES

Symbol Size	Series A Runs and Dev.	Av.	Series B Runs and Dev.	Av.	Total Runs and Dev.	Av.
1/16	201 + 121	5.60	282 + 19	5.07	483 + 140	5.29
1/4	691 + 273	5.40	321 + 51	5.16	1,012 + 324	5.32
1 1/2	576 + 87	5.15	1,197 + 286	5.24	1,773 + 373	5.21
2 1/4			600 + 133	5.22	600 + 133	5.22

runs with the ¼ inch symbols averaged 5.40, and 201 runs with the 1/16 inch symbols gave an average of 5.60. Not only is this difference consistently in the direction of a higher score upon the smaller symbols, but the difference between the ¼ and the 1½ inch sizes has the suggestive C.R. of 2.27 and that between the 1/16 and the 1½ inch ones, the significant C.R. of 2.81. (The smaller number of runs with the 1/16 inch symbols is accounted for by the fact that these were not

introduced until relatively late in Series A.) But as far as the relation between scores and symbol size is concerned, the results of Series B do not follow the earlier pattern. With the 1/16 inch symbols, which had yielded the highest scores in Series A, 282 runs in Series B averaged only 5.07. With the ¼ inch size, 321 runs gave an average of 5.16. The 1½ inch symbols averaged 5.24 for a total of 1,197 runs, and the largest, 2¼ inch symbols, 5.22 for 600 runs. The differences in this series are not as striking statistically as those for Series A. As a consequence of the tendency toward a reversal of results in Series B the combined scores for the total research do not show any effect of size of stimuli upon ESP scores within the limits investigated. This conclusion is indicated by the last three columns of Table III, which show the totals for Series A and Series B combined.

The Relation Between Experience of Subjects and Rate of Scoring

Reference has frequently been made in the literature to a tendency for subjects to decline in scoring ability as they become more experienced with the usual laboratory tests. A consideration of the results of Series A in relation to the previous experience of the subjects with tests of this character led to the suggestion that this phenomenon of declining scoring ability might have produced the differences in results for the various stimulus sizes. This suggestion seemed to offer a possible explanation because of the fact that a larger percentage of experienced subjects participated in the first tests of Series A, when the two larger sizes of symbols were used exclusively, than in the last part when the smallest size was introduced. Therefore, the data of the entire research were analyzed from the point of view of the amount of experience of the subjects to discover whether there was a general tendency of subjects to decline in scoring ability.

For the purposes of this analysis, those subjects who had taken part in any formal ESP tests prior to their participation in Series A were classified for that series as "experienced" subjects. Others in Series A were classed as "inexperienced." In Series B all subjects who had already taken part in Series A or in any other formal ESP tests were classified as "experienced" subjects and all others as "inexperienced" subjects. A subject's classification as to experience was considered to remain unchanged throughout a given series, but the same subject might be inexperienced in Series A and experienced in Series B.

TABLE IV
RESULTS OF SERIES A AND B ACCORDING TO EXPERIENCE OF SUBJECTS

Subjects	Series A Runs and Dev.	Av.	Series B Runs and Dev.	Av.	Total Runs and Dev.	Av.
Experienced	793 + 215	5.27	1,575 + 290	5.18	2,368 + 505	5.21
Inexper.	675 + 266	5.39	825 + 199	5.24	1,500 + 465	5.31

Table IV shows the analysis of the results along these lines for both Series A and Series B separately and for the research as a whole. The slight differences in favor of higher scores for the inexperienced subjects are statistically insignificant.

The Relation Between Newness of Stimulus Sizes and Rate of Scoring

The Problem. However, even a slight difference might lead to the discovery of an important principle. The proposition can logically be formulated in the following manner: All stimulus material used by inexperienced subjects was new to them. On the other hand, only part of the stimulus material used by experienced subjects was new. The difference in favor of the inexperienced subjects might have been caused by the use of a greater preponderance of new material. The problem, then, may be stated: Did stimulus material, when used by a given subject over a period of time, lose its effectiveness as indicated by a falling-off of ESP scores?

The General Evidence. With all the results in hand, it was necessary to set up certain arbitrary criteria of newness of material in order to make a general analysis of the data for a possible effect of a novelty factor. For this purpose, material was classified as "old" for a particular subject in Series A if that size of symbol had been used in formal tests by that subject before he took part in the present investigations. As none of the subjects had worked with any symbols except those of the regular 1½ inch size, the only use of old material in Series A was in those runs by experienced subjects with the regular 1½ inch ESP symbols. All other tests in Series A were considered to be made with "new" material. For Series B, all tests made with any symbol size by subjects who had previously used that particular size of stimulus, either in Series A or in other formal tests before entering upon Series B, were classified as tests with "old" material. All other tests in Series B (including those with all sizes of stimulus for inexperienced subjects and those with sizes used for the first time by experienced subjects) were made with "new" material.

Table V shows the results of the general analysis of the data

into the old and new material categories. The difference between these two groups in Series A is statistically significant (C.R. of the diff. = 3.83) with the higher rate of scoring in the tests with new material. In Series B, a similar division of the work with the four stimulus sizes gives a difference in the same direction, though the rate of scoring with the new material is not significantly higher than that for the old (C.R. of the diff. = 1.50). When both series are combined and the distinction between new and old material is maintained, a significant difference in favor of higher scores with new stimulus material (C.R. of the diff. = 3.43) is obtained for the experiment as a whole.

TABLE V
RESULTS OF SERIES A AND B ACCORDING TO THE NEWNESS OF MATERIAL

Material (Symbol size)	Series A Runs and Dev.	Av.	Series B Runs and Dev.	Av.	Total Runs and Dev.	Av.
Old	331 — 10	4.97	1,493 + 367	5.25	1,238 + 112	5.09
New	1,137 + 491	5.43	907 + 122	5.13	2,630 + 858	5.33
C.R. of diff.	3.83		1.50		3.43	

Further Analyses. The strong suggestion that the sizes of symbols with which the subjects had had less experience were more effective as ESP "objects of perception" raises a number of questions as to the possible nature and origin of this newness factor. Some of these questions can be answered, tentatively at least, by further study of the data from the present investigation. Others can only be stated and considered speculatively. In any event, definite conclusions both as to the actual occurrence of the novelty effect and as to its nature must await further independent experimental confirmation. For what they may be worth, these questions may be raised and considered as far as the results of this investigation will allow.

a. Loss of newness effect. If stimuli lose their effectiveness for ESP scoring with use, the question arises as to when and how the loss occurs. Do the scores with a particular type of symbol drop off gradually, or is there a rather sudden decline after a period of optimum success for each subject? If a point-for-point consideration of the rate of scoring in the present research in relation to the amount of experience of subjects is made, some light might be thrown upon this question. The results of such a study are shown in Table VI, and the same data are represented graphically in Fig. 5.

For the purposes of this analysis, the distinction between Series A

TABLE VI

RELATION OF LEVEL OF SCORING TO THE AMOUNT OF EXPERIENCE WITH THE DIFFERENT SIZES OF STIMULUS MATERIAL

Successive Times of Using (Sessions)	Unclassifiable Runs and Dev.	Unclassifiable Av. Hits per Run	Classifiable Runs and Dev.	Classifiable Av. Hits per Run
1	304 + 17	5.06	1,041 + 409	5.39
2	106 + 0	5.00	571 + 208	5.36
3	44 + 22	5.50	432 + 158	5.37
4	32 + 9	5.28	303 + 6	5.02
5	28 + 7	5.25	240 + 56	5.23
6	52 — 10	4.81	147 + 62	5.42
7	39 + 8	5.21	121 + 22	5.18
8	28 — 11	4.61	75 — 26	4.65
9	19 + 2	5.11	73 + 21	5.29
10	28 + 10	5.36	46 + 9	5.20
11	26 + 14	5.54	27 + 4	5.15
12	17 + 2	5.12	10 — 1	4.90
13	13 — 12	4.08	12 + 3	5.25
14	9 — 8	4.11	11 — 5	4.55
15	14 — 6	4.57		
Total	759 + 44	5.05	3,109 + 926	5.30
	C.R. of diff. = 3.00			

and Series B was disregarded. All tests were divided into two classes. In one, designated as "unclassifiable," was placed all the work *by experienced subjects done with the regular* 1½ *inch symbols, which they had used in tests prior to first starting in the present investigation.* For these subjects, it was not possible to determine the amount of experience with the standard symbols before their participation in this experiment. In the second class, designated as "classifiable," was included the work of all subjects with those sizes which were used only in this investigation. For these tests, the rate of scoring in relation to the novelty of symbol sizes could be traced from the very first session of using any new material through successive occasions of being tested with that same material.

In Table VI all the results of the experiment are presented as they belong under these two groups. The results for all symbol sizes from the first experimental session in which each was used in this research are brought together as Session 1. The results of the second occasion of using each size are combined as Session 2, and so on for the entire experiment. The smaller numbers of runs for later sessions are due to the fact that not all subjects used each symbol size equally often and that the subjects served for an unequal number of sessions.

An examination of the unclassifiable column shows no significant trend in scoring throughout successive periods of working. As we

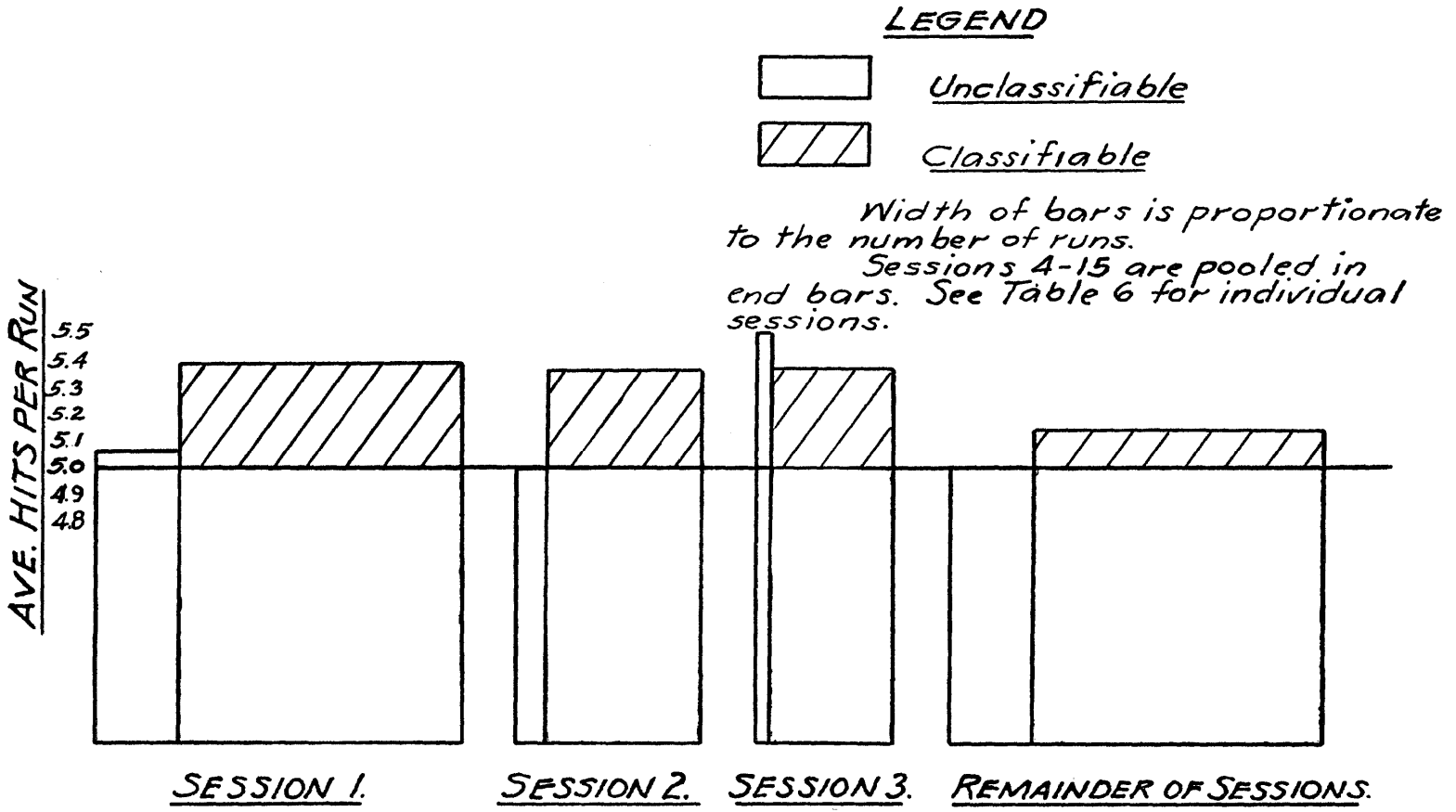
LEGEND
Unclassifiable
Classifiable
Width of bars is proportionate to the number of runs.
Sessions 4-15 are pooled in end bars. See Table 6 for individual sessions.
AVE. HITS PER RUN
5.5
5.4
5.3
5.2
5.1
5.0
4.9
4.8
SESSION 1.
SESSION 2.
SESSION 3.
REMAINDER OF SESSIONS.

Figure 5

should expect from the previous indication of the favorable effect of novelty, the results for this classification of symbols, with which the subjects had had an indeterminate amount of experience before the first session with them in this investigation, were low in general rate of scoring (an average of 5.05 for 759 runs).

On the other hand, the classifiable column, in which the subjects' experience with the various sizes of symbols can be traced, session by session, from the very first use of each size, seems to tell a different story. The first three sessions with new material give a fairly uniformly high average. Thereafter, fluctuations in average from session to session appear in a manner suggestive of those of the unclassifiable column from the start. If all the sessions in the classifiable column from 4 to 14 inclusive are combined, the average is found to be 5.14 for 1,065 runs. As contrasted with the average of 5.39 for the 1,041 runs of the first session and that of 5.36 for the 1,003 runs of the second and third sessions combined, this suggests that the subjects tended to obtain lower scores with a particular new size of stimulus symbol sometime after they had used it for three experimental sessions. No generalization to other experiments is possible and none is intended. In connection with the general evidence that novelty favors ESP scoring, the data of Table VI and Figure 5 are important in that they show an actual decline did occur for the successive uses of sizes of stimuli to which the subjects were introduced for the first time in this experiment.

b. Relation between experience and effect of newness. The question arises as to how experienced and inexperienced subjects compared in their ESP performance on new material. Did subjects who were experienced on old material before a given series and who got lower scores upon the old material, do worse with new material than the inexperienced subjects, for whom all material was new? The data in Table VII clearly indicate that new material was equally effective for experienced and inexperienced subjects.

TABLE VII

A COMPARISON OF "EXPERIENCED" AND "INEXPERIENCED" SUBJECTS WITH "NEW" MATERIAL

Subjects	Series A Runs and Dev.	Av.	Series B Runs and Dev.	Av.	Total Runs and Dev.	Av.
Experienced	462 + 225	5.49	668 + 168	5.25	1,130 + 393	5.35
Inexper.	675 + 266	5.39	825 + 199	5.24	1,500 + 465	5.31

c. Relation of subject's knowledge of material to effect of newness. It will be recalled that different methods were used for determining the order of presenting symbol sizes within an experimental session in which two or more sizes were used. An analysis distinguishing among the various methods of selection (regular alternation or rotation in which the subject knew when each size was to be used; or the experimenter's subjective determination of the order or following the cast of a die in which the size was not known to the subject until the end of the run) showed no significant differences among them for the general results. However, a question arises in connection with the effect of novelty of sizes and the fact that the subjects sometimes did not know during a run what size stimuli were being used: namely, what was the relation between the effect of new material upon scoring and the subject's knowledge of the kind of material being used?

In other words, the analysis presented in Table V showed that, in general, subjects scored better with new than with old material. The question now raised is this: Did that relation hold both when the subjects knew and when they did not know what size of stimuli was being used? Table VIII shows that the difference in scoring rate in favor of the new material was much greater when the subjects knew before each run what stimuli were to be used than when they did not know. Indeed, the difference in averages is statistically significant for the tests in which they knew about size of the stimuli (C. R. of the diff. $= 3.75$), which is not the case for those tests in which they did not know until after the run with which size of stimuli they were being tested (C. R. of the diff. $= 1.09$). However, when the subjects were kept ignorant of the size of the stimuli, the average with the old material was slightly higher and that of the new material lower than the general averages for each (see Table V). Consequently, the general average for all tests in which subjects did not know the size of symbols during the run is insignificantly below that of the total results of tests in which they knew.

TABLE VIII

A COMPARISON OF "NEW" AND "OLD" MATERIAL WHEN THE SUBJECTS KNEW AND DID NOT KNOW WHICH STIMULI WERE BEING USED

Classification of Stimuli	Subjects not Knowing Size		Subjects Knowing Size	
	Runs and Dev.	Av.	Runs and Dev.	Av.
Old	549 + 83	5.15	689 + 29	5.04
New	606 + 169	5.28	2,024 + 689	5.34
Total	1,155 + 252	5.22	2,713 + 718	5.26

d. Scoring trends within experimental sessions. A point of interest in relation to Table VI is that of how the run-by-run performance curve of the new (classifiable) material compared with that of the old (unclassifiable) material. One question has to do with the trend of the scores during the *first few runs of the session*—or of Session 1 in particular. Rhine (11) reported a characteristic period of adjustment to a new condition reflected in the scores by a rising level of scoring during the first few runs. Another question is that of how the performance with the new and old material compared throughout the experimental session.

In Fig. 6 the results of the first two sessions of Table VI are shown graphically. The curves suggest that there was an adjustment period in the first session with both old and new material. Also, the average difference in favor of the "new" material seems to have resulted from a more consistent level of scoring. Because of the fact that sessions for various subjects were not of equal length, the points on the curves toward the end of each session do not represent as many runs as those toward the beginning. Each line of evidence, however, is only suggestive in character and takes on real significance only if compared with other similar lines. The evidence of an adjustment period may therefore be said to be stronger because it confirms the observations of earlier investigators, while the suggestion of a steadier, more sustained, rate of scoring upon new material remains to be confirmed or refuted by subsequent investigations.

DISCUSSION

In general, the evidence on the relation between the kind of stimulus material and the rate of scoring shows that some characteristic or characteristics of the symbol material affect the degree of success in identifying the symbols. The foregoing analyses point to the newness of stimuli as the most important factor. It is appropriate to inquire whether this was the *only* factor and to discuss how the effect of newness upon ESP scoring is to be interpreted.

There is more than a superficial relation in the results between size of stimuli and newness. The fact that the rate of scoring in Series A was inversely related to the size of stimuli was logically interpretable as due to either factor, size or newness. However, there would appear to be no basis for expecting the effect of size, if any occurs, to be in an *inverse* relation to the scoring level. This fact, in itself, strongly suggested that the explanation lay either in the new-

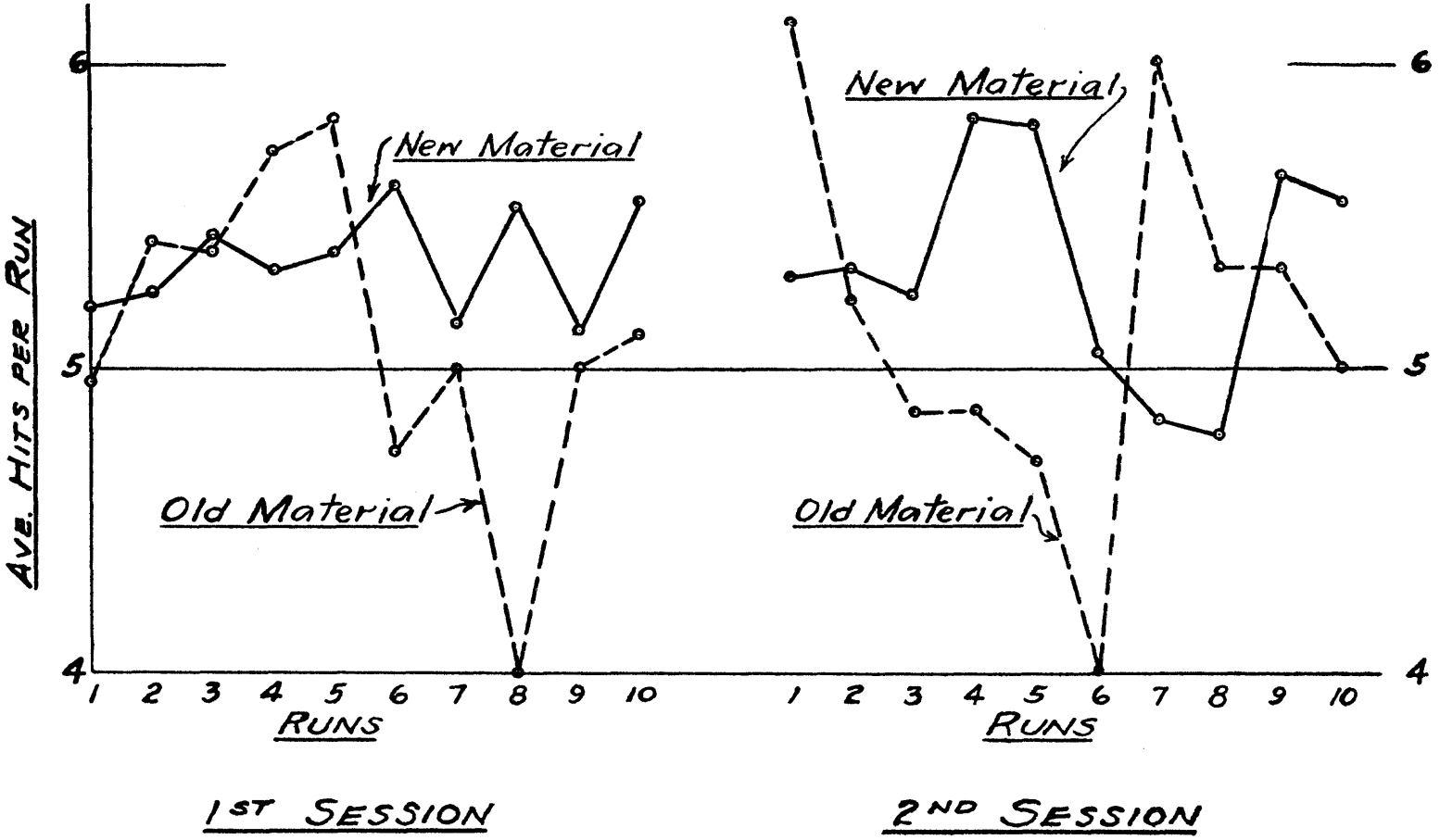
AVE. HITS PER RUN
6
5
4
New Material
Old Material
1 2 3 4 5 6 7 8 9 10
RUNS
1ST SESSION
New Material
Old Material
1 2 3 4 5 6 7 8 9 10
RUNS
2ND SESSION

Figure 6

ness factor or in still another one. It has been made clear from the presentation of the data that this suggestion was supported by Series B and the combined results of the entire experiment. The results from this investigation must therefore be taken as strengthening the general evidence that the known physical characteristics of the stimuli do not affect ESP. Size was not a factor within the limits of variation introduced in this study.

The factors of experience of subjects and newness of material are obviously closely related. In ESP tests in which the same material is used throughout, it is not possible to distinguish between the general experience of the subject and his use of the standard ESP symbols, as they both increase together. The present investigation involved variations in conditions such as to permit an analysis of the data distinguishing between these two factors.

The evidence here points strongly to a direct relation between the level of scoring and the number of runs for which each size of symbol had been used. This relationship is one which may offer further explanation of the generalization made by other investigators (see p. 123) that scoring ability decreases with experience. Subjects did decline with experience in this study as well, but this effect tended to be specific to each size of stimulus rather than generalized. Subjects who had used the standard ESP symbols before entering upon tests with new sizes did just as well with the new symbols as did completely naive subjects.

An important implication of the results is the suggestion that the scoring rate was kept at about the same level throughout the experiment by the introduction of new material when, presumably, it would have declined much sooner if the same size of stimuli had been used throughout. It is not yet possible to generalize for these subjects as to whether they can be made to score above expectation for an indefinite period simply by introducing new material at strategic points. Nor is it possible to state whether changes other than those involving size would affect the results in the same way. There has been a general clinical impression abroad among ESP workers that a change of conditions helps to keep the subject interested in the tests in a way that favors scoring. Further direct experimental evidence to define the conditions under which this generalization is applicable is obviously needed.

Unfortunately, the quantitative results of the research do not point the way to a definite interpretation of the psychological difference

between new and old material. General observations suggest, however, that the experimental situation was such as to elicit more favorable motivation in the subjects when the new stimuli were used. One possible interpretation would be that the effect is related to general psychological satiation for the various kinds of material. If this were true, the decline in scores with the use of a particular size of symbol might be considered to be a psychological aspect of the relation between the subject and the stimulus, and the changing of other features of the experimental conditions might not have the same effect as changing the stimuli.

On the other hand, there are indications that the newness effect is partly, at least, a function of the relation between the experimenters and the subject; or, in other words, that the experimenters unintentionally used the new sizes of symbols as an opportunity for eliciting a more favorable attitude in the subjects. The experimenters always showed the subjects the new sizes before beginning the daily session in which they were first used. This was frequently done in a manner which challenged the subject to do better with the new sizes before beginning the daily session in which they were first used. During experimental sessions the experimenters, particularly Woodruff, frequently encouraged the subject between runs in the same challenging manner. The introduction of a new size of stimulus was made a special "talking point" between the experimenters and the subject. This probably resulted in a greater interest of the subjects in the new material for a few sessions, after which the challenge to do better either lost its effect or was shifted by the introduction of another new size. These speculations serve chiefly to emphasize the need for further research.

SUMMARY

(1) The results of Series B appear to bear in a crucial manner upon the problem of the occurrence of ESP. In that part of the research the conditions were carefully planned to control against the effects of hypothetical sources of error by explicit steps in the experimental procedure. These safeguarding conditions have already been summarized on p. 126 ff. The results of this period of joint investigation included 2,400 runs with a deviation of 489 hits beyond mean chance expectation, an average of 5.20 hits per run. The S.D. for 2,400 runs is 97.98, and the C.R. of the observed result is 4.99. The P-value for the result is 3×10^{-7}; when allowance is made for the

possible effect of optional stopping, this is increased to $P = 5 \times 10^{-6}$. The conclusion was reached that "perception without the use of recognized sensory channels" is the only principle which can reasonably account for the results of Series B.

Because of the essential similarity in results between the two main divisions of the research, this conclusion was extended—for purposes of further analysis of the data as they bear upon the nature of ESP—to the results of Series A as well.

(2) Analysis of the data according to the size of stimulus symbols used showed that size *per se* did not affect the results of the investigation as a whole (Table III). Suggestive differences among the sizes used in Series A in favor of the smaller stimuli proved on further analysis to be related to the newness of testing material.

(3) The amount of experience of subjects with ESP tests was not, in itself, directly associated with trends in scoring (Table IV). Again, there were suggestive differences in the direction of higher average scores by the inexperienced subjects, but these also proved to be related to the newness of testing material.

(4) When the results were considered in relation to the amount of subjects' experience with the different kinds (sizes) of stimulus material, subjects were found to score significantly better with new than with old material (Table V). The advantage of working with new material was found to decline after a time as the subjects became more experienced with the new material (Table VI). Experienced and inexperienced subjects scored equally well with new material (Table VII). When subjects did not know before or during a run what size of stimuli were used, the advantage in favor of higher scores with new material was not statistically significant (Table VIII). The evidence suggests that there was a period of adjustment during the first runs in the first sessions of using any material, whether new or old (Fig. 6).

REFERENCES

1. Carpenter, C. R., and Phalen, H. R. An experiment in card guessing. *J. Parapsychol.*, 1937, I, 31-41.
2. Dunlap, K., and others Adequate experimental testing of "extra-sensory perception" based on card sorting. *J. Parapsychol.*, 1939, III, 29-37.
3. Gibson, E. P. A study of comparative performances in several ESP procedures. *J. Parapsychol.*, I, 26-275.
4. Greenwood, J. A. Analysis of a large chance control series of ESP data.

J. Parapsychol., 1938, II, 138-146.
5. Greenwood, J. A., and Stuart, C. E. The mathematical techniques used in ESP research. *J. Parapsychol.*, 1937, I, 206-225.
6. Leuba, C. An experiment to test the role of chance in ESP research. *J. Parapsychol.*, 1938, II, 217-221.
7. MacFarland, J. D., and George, R. W. Extra-sensory perception of normal and distored symbols. *J. Parapsychol.*, I, 93-101.
8. Murphy, Gardner, and Taves, Ernest Covariance methods in the comparison of extra-sensory tasks. *J. Parapsychol.*, 1939, III, 38-78.
9. Pratt, J. G. Clairvoyant blind matching. *J. Parapsychol.*, 1937, I, 10-17.
10. Price, M. M., and Pegram, M. H. Extra-sensory perception among the blind. *J. Parapsychol.*, 1937, I, 143-155.
11. Rhine, J. B. *Extra-Sensory Perception.* Boston: Boston Society for Psychic Research, 1934.
12. Rhine, L. E. Some stimulus variations in extra-sensory perception with child subjects. *J. Parapsychol.*, 1937, I, 102-113.
13. Riess, B. F. A case of high scores in card guessing at a distance. *J. Parapsychol.*, 1937, I, 260-263.