APPRAISING VERBAL TEST MATERIAL IN PARAPSYCHOLOGY

By J. G. Pratt and William R. Birge

ABSTRACT: The writers trace the efforts that have been made toward finding a satisfactory means of assessing the accuracy of verbal test responses. The descriptive statements which make up this material appear in many forms: as mediumistic utterances, as impressions given in association with a personal object, as automatic writing, and the like. Progress toward a method is described step by step, beginning with the work of Saltmarsh in 1930. The emphasis is placed upon the advance represented by each new contribution. The current procedure is explained, and suggestions are made for further research.—Ed.

Introduction

This will chiefly be an account of steps that have been taken in the Parapsychology Laboratory over a period of fifteen years toward the development of a method of evaluating verbal material obtained in tests of psi abilities. It will be a report on a search for a method, a search that has resulted in definite progress even though it is still not complete. Thus far in working with responses given in the form of descriptive statements, the investigators have been trying mainly to find a way of appraising the results, and relatively little attention has been paid to obtaining the best possible subjects for such psi tests and to providing the most favorable psychological conditions.

Verbal material is one of the forms in which ESP has been most frequently reported to occur. If an adequate method can be found, a careful study of this type of response obtained under suitable test conditions would obviously have wide application and great value. The type of word material that has received the most attention in parapsychology has been the utterances of "mediums." These statements have been studied chiefly from the point of view of their bearing upon the hypothesis of spirit agency. The study of mediumship depends largely on a method for accurately evaluating verbal material, and we may say that progress in such study has long awaited the development of methods such as we are seeking.

Early Efforts at Appraisal

In spite of the long history of the investigation of mediumistic records, the first efforts to find a way of appraising verbal material were made less than two decades ago. The pioneer in this line was H. F. Saltmarsh, who published the first report on this topic in 1930 (3).

Saltmarsh used as the subject of his experiments the English medium, Mrs. Warren Elliott. Sometimes the person for whom the utterances were intended—that is, the "cooperator" in a test—was present at the time the subject gave her responses, but at other times the tests were made with the cooperator absent. In the latter case, the subject was handed some small personal belonging and was asked to give her impressions regarding the circumstances of the people who had been closely associated with the object. It was these "token object" tests that Saltmarsh used in trying out new methods of evaluation.

In the method of scoring that Saltmarsh first proposed, the procedure was as follows: To start with, he broke the subject's verbal material down into separate items or statements. He then submitted each of these itemized records to the cooperator to be annotated regarding the applicability of the statements to either his own personal circumstances or those of the deceased owner of the token object. At the same time, he submitted each record to a "control" cooperator, asking him to comment on the items as if the information were intended for him. On the basis of the comments made, Saltmarsh decided whether each item was true or false for the cooperator and for the control annotator.

This procedure for marking the records was a distinct advance over anything done up to that time, but as a pioneering effort it naturally had some weaknesses. The most serious of these was the fact that the cooperators knew they were commenting upon material that was intended for them, and there was no way of knowing how this knowledge affected their scoring of the items. Also, the control markers knew that the information was *not* directed toward them, and this fact introduces a real doubt as to whether their judgments were comparable with those of the cooperators. Finally, Saltmarsh himself knew which comments were made by the cooperators and

which ones by the control markers. It is possible that his judgments reflected some degree of subjective bias in deciding in each instance whether an item was true or false.

The method involved several further steps before a final assessment of the material was reached. Saltmarsh classified the items of each record in terms of three groups: clichés, which would be widely applicable; definite statements, which would apply less widely; and characteristic statements, which would be true of very few people. To these he assigned values of 1, 5, and 20, respectively, for each item, and on the basis of these values he then gave two scores to each of the subject's test records, one for the cooperator and another for the control annotator. The criterion of significance arbitrarily set up was that the cooperator's score on his own record should be at least eight times as great as the score obtained by the control marker.

In the second attempt to improve the methods of appraising verbal test material Saltmarsh was assisted by S. G. Soal (4). This study made use of a more precise statistical procedure for measuring the significance of the results. The basic mathematical method was suggested by Soal and further improved by R. A. Fisher. It provided a means of evaluating a series of items regarding each of which two things were "known": (1) whether each was *true* or *false* for the actual cooperator; (2) the value of its probability ratio, or its likelihood of being true of anyone selected at random from Western adult civilization. (Saltmarsh arrived at the "truth" or "falsity" of those statements that were relevant to each cooperator, as he had earlier, by having each one annotate his own material and then by deciding himself whether or not the item applied. He also assigned the item probability ratios by personal judgment.) Given these two sets of facts, the method made possible an assessment of the results in terms of the probability that the results would by pure chance have fitted the cooperators as well as they did.

As we see it now, there was a fundamental weakness in this proposed statistical analysis that was caused by the nature of the material to which the method was to be applied. The Saltmarsh-Soal formulas assumed that the items to be evaluated, or the data derived from them, were independent of one another. There is no

way of knowing that this assumption is justified when the items are drawn from general descriptive accounts. For example, the subject in a token object test of his psi abilities might start out to talk about an English cottage. Having chosen this theme, he might go on to give several items appropriate to it. If an English cottage happened to have some personal significance for the cooperator involved, many of the items might be correct simply because the description was internally consistent. On the other hand, if the general topic of the description happened to be wrong, all the items might be rejected for this reason. Either way, the items would not yield statistically independent data for analysis.

Quite apart from the interdependence of the items themselves, this difficulty in the way of statistical assessment might arise from subjective factors of judgment on the part of those who score the items. If a person knows that the information he is asked to check is intended to fit his circumstances, he may score it differently than he would if he thought that it was not his. This emphasizes the importance of having the cooperators ignorant of whether or not the material they are asked to mark is their own.

A further weakness in the new method as it was first applied lay in the fact that Saltmarsh arbitrarily decided the item probability ratios on which the evaluation was based. That is, he used subjective judgment in deciding how likely it was that a given item would be true purely by chance. In making a trial application of the method to one of Mrs. Elliott's records, he pointed out that it would have been better if the average of several independent judgments had been used.

The work of J. F. Thomas (5) represented the first large-scale attempt to apply the Saltmarsh-Soal method of scoring verbal material. Making the advance in methodology that the earlier workers had suggested, he averaged the opinions of several judges to get the probability value of each item. He was the person (cooperator) for whom all the test statements were intended, and he scored them all as being correct, incorrect, inconclusive, or unverifiable. Only the first two classes, the "true" and "false" items, entered into the final analysis. Two separate committees of judges were asked to assign probability estimates for these items, and duplicate evaluations

were worked out on the basis of these two sets of ratios. In one case, a critical ratio of 13.04 was obtained; in the other, 4.91. Each of these is significant, but the wide discrepancy between them suggested to Thomas a weakness in the method of using estimated probability ratios.

A second statistical treatment applied to these records was that of having a number of "control" cooperators each check the material for applicability to himself, just as Thomas had done. By an appropriate method of analysis Thomas' score was then compared with the distribution of scores made by the control group and a measure of its significance in terms of this general background was obtained.

Thomas' methods of assessing the subject's success in giving correct information suffered throughout from the same basic weaknesses as the earlier efforts. He did not achieve objective judgments of the material, as everyone who scored it knew whether or not it was intended for him. Also, the use of the Saltmarsh-Soal formulas here, as in the original study, involved the unwarranted assumption that the items within each record, or the data derived from the items, were statistically independent.

The first efforts to advance the methods of appraising verbal test material made in the Parapsychology Laboratory were based upon experiments in which Mrs. Eileen Garrett was the subject. There were two separate series of tests, one in the spring of 1934 and the other a year later, and a full report of this work was published (2). In the present paper we shall be concerned only with reviewing briefly the advance in methodology that each series represented.

In the first series of tests with Mrs. Garrett a new method of deriving the item probability ratios for use in the Saltmarsh-Soal formulas was introduced. In this experiment there were twelve cooperators, and one session was held for each, with the cooperators present. The twelve records were later broken down into separate items by the experimenter and all of them were submitted to all of the cooperators, who were asked to check each item as correct (√), incorrect (X), unknown (?), or not relevant (in this case the space within the parentheses was to be left blank). Only those items from each record which were marked either as correct or incorrect

by the actual cooperator were used in the final evaluation. The probability ratio of each of these items was obtained from the way the other eleven cooperators checked each one. The denominator of the ratio was taken as the combined number of those marking the item as correct or incorrect, while the numerator was the total of those checking it as correct. When the actual cooperator was the only one who checked an item as correct, his marking of the item was included with the other eleven ratings in getting the probability ratio. This was done to avoid having an item probability of zero, which would occur when the statement was true for no one except the actual cooperator.

The advance represented by the first series, therefore, was that of having the cooperators themselves provide the control checking of the items. In this way it provided empirical item probability ratios and met the objection offered by Thomas to the earlier attempts to use the Saltmarsh-Soal method.

The second series of tests with Mrs. Garrett made a still further advance toward an adequate method of assessing verbal material. It introduced conditions to keep the cooperators unaware of which were their own records until after they had marked them all. Fifteen cooperators took part in this experiment. At the time of each test, the cooperator concerned was kept in a room adjoining that of the subject, with the intervening door closed. The identity of the cooperators was withheld from the subject, who worked in this series, as she had in the former one, in a trance state. Each cooperator furnished a token object. The subject knew that these were available, but she did not always use them.

After the fifteen tests had been conducted, the experimenter analyzed each record into items or points as in the first series. As each one was prepared it was mailed out to the cooperators to be checked. Each cooperator was asked to mark all the material as if it applied to himself. Since none of the cooperators was in the room with the subject at any of the tests, they were not able to identify their own records from memory. This control of the subjective factor that was present when the cooperator knew which record was his own was the outstanding advance in method achieved in this series of tests.

From the scored records, the item probability ratios were obtained as they had been in the first series. These were then substituted in the Saltmarsh-Soal formulas to arrive at an evaluation of the results. The evaluation showed that three of the fifteen subjects who took part in the second series obtained an individually significant rating on their own records, and that the fifteen records as a whole were highly significant.

Was this an adequate method for testing the significance of these results? The original report discussed one possible loophole in the procedure. Some of the cooperators might have obtained significant scores on their individual records because they consistently accepted a large percentage of items throughout the entire fifteen records. To check up on this possibility, a study was made of the "yes"-tendency in the three cooperators whose records turned out to be significant. In two instances there was no general tendency to accept a large percentage of items in all the records, but there was in the third case. In other words, the results of two significant tests were not attributable to this factor. It was pointed out, however, that even such a general "yes"-tendency as was exhibited by the third cooperator would have the effect of *lowering* the significance of the other records and thus the exaggerated value given his own material would be offset to some degree when the series was evaluated as a whole.

However, there is still a serious difficulty that prevents our concluding that the subject demonstrated psi ability in these tests. This is the matter of the interdependence of the utterances given in each session, the characteristic of descriptive statements that was explained in connection with the first application of the Saltmarsh-Soal method. It is a weakness that appears to be inherent in the method itself, a fact that has not always been as fully appreciated as it is today. Finding some means of overcoming this difficulty has become one of the primary objectives of the efforts to devise a method of assessing connected statements.

Methods Tried More Recently

Since the publication of the preliminary report on the assessment of mediumistic utterances in 1936, the Parapsychology Laboratory has continued the efforts to improve the methods for dealing with

verbal test material. During this period we have worked almost exclusively with the results obtained in token object tests. We used a number of subjects who thought they might be able to demonstrate their psi abilities in this way. The objects were always securely wrapped and sealed to avoid the possibility of inferring anything from them. In most instances, the cooperators submitted objects that were keepsakes of high sentimental value because of their connection with someone deceased. In this way we were able to work more effectively with subjects who in their experiences along these lines had been oriented toward the spiritistic hypothesis.

As we stated earlier, the emphasis thus far has been kept on the development of a suitable way of appraising verbal test material, and we shall continue to limit ourselves in this paper to this question. The results, which have not been significant as a whole, will be presented only as needed to illustrate a method. Nor shall we be concerned at this time with the details of procedure in these token object tests that do not directly affect the assessment technique. Suffice it to say that precautions have been taken to insure that the subjects should not obtain any clues regarding the identity of the cooperators and the original owners of the token objects.

In addition to the shortcomings discussed earlier, all the methods already described suffered from the practical difficulty of being cumbersome to apply. This difficulty became especially great in the methods that were first tried in the Parapsychology Laboratory. The use of control checking of the material by the cooperators led to the necessity of treating a larger number of test records at a time in order to increase the reliability of the item probability ratios. As work along these lines was resumed, it was with a clearer recognition of the fact that the method of assessment would have to be one that investigators would find practicable.

Another practical difficulty was that some of the methods had been wasteful of the subject's utterances themselves. A literal interpretation of the statements had been followed, and as a consequence many of the items were excluded from the final evaluation as having no relevance for the cooperator for whom they were intended.

In 1944 C. E. Stuart proposed a method that seemed to overcome these two practical difficulties to some degree. One suggestion

he made was that when the records were broken down into items the wording used should be such that each cooperator would be able to mark a larger percentage of the statements as either correct or incorrect for himself. Stuart took the view that the subject might be giving correct information in his utterances, but that he might be directing his remarks less accurately than a literal interpretation would suggest. Such things as mistaken identity or the mixing up of two or more individuals connected with the cooperator might be occurring. Thus, suppose the subject said of a particular cooperator that he had a deceased uncle and that this uncle had been fond of smoking a corn-cob pipe. According to the earlier methods of itemizing the material, the descriptive item about the corn-cob pipe would necessarily be considered of no relevance to the cooperator if he did not have a deceased uncle. With Stuart's method, the item might still be considered applicable provided the cooperator had a deceased father or some other relative who fitted the description. For example, this particular item might be presented in two parts, in some such manner as follows: "You have a deceased relative who used to be fond of smoking a corn-cob pipe ( )." "This relative was your uncle ( )." This broader basis for the interpretation of the material proposed by Stuart was in keeping with the view that the subject's verbal responses might be even more "free" than the actual words of the subject seemed to imply.

Stuart also suggested a method of evaluation that required fewer cooperators for an experiment. He proposed five token object tests as the standard number. An ingenious evaluative procedure was devised that still made use of empirical ratios but did so with only five cooperators. As in the earlier work, each cooperator was asked to score all the records without knowing which one was his own. Stuart selected for evaluation only those items to which four cooperators had responded with "no" and one had responded with "yes." On a purely chance basis, each of these items has a one-fifth probability that the person who scored the item "yes" would be the one for whom it was intended. If the cooperator who scored the the item "yes" was the same cooperator for whom the subject intended that item, it was counted a hit. Otherwise it was a miss. The statistical significance of a particular record as well as of a series

of five records could then be computed in exactly the same manner as that in which the deviation for a particular number of trials in a standard ESP card test is measured.

Stuart had not made a final formulation of his method before his untimely death in March of 1947. His procedure as described here offered definite advantages of the sort that he was trying to achieve. There were still some shortcomings in the procedure, however, as Stuart himself was well aware. One of these was the fact that the method, in spite of achieving a much higher percentage of judgments of the material from the cooperators, still allowed only a small amount of the information given by the subject to enter into the final calculation of the results. Another objection, one of an even more fundamental character, was that the statistical evaluation used still involved the assumption that the items were independent of one another. This requirement concerning item independence was not met in this procedure any more than it had been in the methods already described.

Stuart also made an effort to apply to verbal material the evaluative technique which he had developed for free response tests based upon pictures as targets. This is the preferential matching method that is now widely used in drawings tests. There seemed to be no a priori reason why this method should not be used for free verbal material.

In applying this method, Stuart broke down each record in a set of five token object readings into separate paragraphs dealing with distinct topics or personalities. The records were then sent to the cooperators in units of five paragraphs, including one paragraph from each of the original five records. The cooperators were asked to rank each paragraph in a unit according to its applicability to their circumstances. This method was later discarded as too wasteful, since it lumped together all the information in a paragraph as a single trial by the subject.

After Stuart's death, one of the writers (W.R.B.) for a time assumed an active role in the effort to improve the methods for handling free verbal material. Still another new approach was tried. As before, the information in a set of five records was itemized and submitted to the five cooperators without their knowing which were

their own records. For the evaluation of the results, only those items in all the material that were checked by each cooperator were counted as "trials" made by him. Of these trials, those check marks which came within his own record were considered as hits. If the records were of the same length, in the sense that each one offered the same number of items to be accepted or rejected, the expectation was that the number of check marks made by a cooperator within his *own* record would be one-fifth of the total number of his check marks in *all five* records. Any observed tendency for a cooperator to check a larger number of items correct within his own record could be evaluated in terms of the deviation from the expected number by the same formula as Stuart had used—the one that is commonly applied in ESP card work.

This method did seem to represent a definite step forward. It was a more simple and direct approach to the problem than the procedure Stuart had proposed, and it also enabled the investigator to make use of all the items that the cooperators checked as correct for themselves. Its shortcomings were that: (1) like the earlier procedures, it assumed an independence of items; (2) records of different lengths presented a special difficulty in that they could not be used without changing the 1/5 probability basis of evaluation required for the statistical analysis. The practice proposed was that of itemizing all the records and then dropping off items at the end so that all the records would have the same number of items as the shortest one. This device again involved the wasteful necessity of leaving out some of the material.

The Current Procedure

Up until this time all of the methods that had been tried, except that of treating paragraphs as a whole by means of the preferential matching method, involved editing of the free verbal material, or breaking it up into items before it was scored by the cooperators. The practice in most of the methods of assessment tried had been to separate out the items and to do some paraphrasing of the subject's utterances in order to help the cooperators in their marking of the material. The idea occurred to one of us (J.G.P.) that keeping the exact words used by the subject would offer several distinct advantages. In the first place, this procedure would save a great

deal of time. Furthermore, it would avoid the danger of misinterpreting the verbal material in a manner that might affect the scoring. This danger, of course, becomes serious only in case the person who itemizes the material is acquainted with the cooperators. It seemed that it might be best to avoid, if possible, all "editing" of the records. Experience with these records suggested that it might be possible to achieve all the advantages of itemization simply by inserting parentheses ( ) wherever the subject made a remark that introduced a new thought or qualified a statement in any way. The only judgment required was in deciding where to insert one of these checking points in the material.

All of the foregoing methods for evaluating verbal test material in terms of separate items assumed an independence among the items that may not have existed. For the items within a record to be truly independent, the fact that a cooperator has checked a particular item as correct or incorrect should have no relationship to the manner in which he checks the remaining items in the record. There are two main reasons why the items cannot be considered independent in this sense. First of all, descriptive accounts tend toward some degree of self-consistency: items pertaining to the same topic are likely to be highly related. It follows that, if a particular item is correct or incorrect for a cooperator, an indeterminate number of other items are likely to be checked as correct or incorrect purely as a consequence of this interrelationship or self-consistency.

The second factor undermining the assumption of independence of items is the fact that, quite apart from the self-consistency of a record, the answers made by the cooperators may not be independent of one another. Thus a cooperator may form the general opinion that a record is or is not meant for him on the basis of a few introductory items. Through a perseverative tendency, the cooperator's scoring of the remaining items in the record may show a general effect of this opinion.

The method currently being tried out in the Parapsychology Laboratory does not assume independence of items. The logic of this new method may be explained by an illustration drawn from familiar areas of ESP research. Investigators have long been aware that data obtained from group tests may not be interpretable

in terms of the same statistical analyses as apply to the data of individual tests. In group tests, a number of subjects attempt to call the same targets, and there is no way to be sure that any nonrandomness found in the calls made by different members of the group could be attributed only to the special ability (ESP) being tested. For example, subjects in a group might exhibit a tendency to start their calls with a certain symbol, or they might show the same symbol preference in their series of calls because of cultural or environmental influences or for some other reason having nothing to do with ESP ability. Any such group pattern or characteristic of response would not change the expected average number of hits, but it would affect the variance, the distribution of total scores about the mean for the group. The data could not properly be evaluated on the binomial hypothesis, which assumes that the calls are independent. For the variance of the binomial hypothesis to apply, it is necessary that at least one of the series of events to be compared—either the cards or the calls—be random. This condition holds in a test in which a single subject calls a particular card order only once; for the test to continue, a new card order is provided. In this case the use of a new order of target symbols for each run meets the requirement for randomness. The difficulty in the group situation is that the same card order is used over and over again in checking the calls of the individual subjects.

What statistical procedure might be used to evaluate group tests without making the assumption of independence of results when many call sequences are compared with the same card order? Greville worked out and published a solution to this problem (1). His procedure involves taking the actual distribution of calls in the separate trial positions (the calls actually made on each target symbol) as the given data. For example, consider the case of 100 subjects attempting to call through a single random order of 25 ESP symbols. The results would be evaluated in terms of 25 trials, not as 2,500, the actual number of calls made. The statistical question asked is as follows: Given the particular distribution of symbols in the calls as observed in each of the 25 trial positions, what is the probability that this fixed distribution would give a total score as high as the one made by the group? The total score on the test

would be found by comparing the distribution of calls with the order of symbols in the target deck, this order being a random arrangement which is only one of a large number of possible permutations of the 25 symbols. For each of these permutations of the target deck, a score could be obtained by checking the order of symbols against the "fixed" or observed distribution of the subjects' calls. If the deck were arranged in every possible way and the scores were checked for every permutation, it would be possible in theory, though impracticable in fact, to work out the mean score and the variance of the distribution of all these scores. The score made by the group in the actual test (that is, the hits obtained when the observed distribution of calls is checked against the particular random order of 25 card symbols used) could then be measured in terms of its deviation from the mean and the standard deviation of the entire distribution.

Greville developed the formulas for applying this particular test of significance without the necessity of actually permuting the target order through all its possible arrangements. He considered two general situations: first, that in which the target order represents a truly random selection from among the choice possibilities offered; secondly, that in which a closed deck is used, presenting an equal number of all the possibilities with a random arrangement such as might be given by adequate shuffling. Only the closed deck situation is relevant to the problem of evaluating verbal material.

In thinking over the difficulty which the interdependence of items in long descriptive accounts seemed to cause for statistical assessment, it occurred to one of the writers (J.G.P.) that the Greville method for evaluating the data of group card tests could also be applied to the results obtained from token object tests. As a means of making this application easier to follow, we shall first give another illustration of the use of the method in evaluating the results of a group test involving card calling. In this instance we shall set up a hypothetical situation that is more closely parallel to that represented by token object tests.

Assume that a group of 100 subjects is being tested for ESP ability. They are told that the test will consist of five trials. For targets, cards bearing the common surnames Jones, Brown, Smith,

Hill, and Greene will be used, each target being used only one time during the five trials. The target order is thus one of the 120 possible permutations of the five target names.

Assume, further, that the instructions are given that each subject is to make a response on any particular trial only if he feels confident of making a hit. This would lead to differences in the numbers of calls from trial to trial, a condition which presents no difficulty in the use of the Greville method of analysis. The method, in other words, makes possible the computation of the mean and the variance of the scores that make up the general distribution to which the actual score belongs regardless of unequal numbers of calls from trial to trial. Those who are interested in the technical statistical aspects of this method will see the reason for this from Greville's original article. Others may accept it as a statement of fact.

Assume that when the data are tabulated in terms of the frequency of calling each name on each trial, the following distribution of responses on the five trials is found:

Trial	Subjects' Responses				
	Jones	Brown	Smith	Hill	Greene
First	9	7	0	0	3
Second	3	0	0	0	12
Third	10	0	26	3	13
Fourth	2	0	0	2	12
Fifth	3	0	0	4	4

The Greville method takes these figures as the given data, and it thus avoids making any assumptions regarding statistical independence among these observed responses. Taking these data as they are, the Greville method enables us to find the *mean* and the *variance* of the 120 scores that would be obtained when this particular matrix of responses is checked against the 120 different orders in which the five target names might have been presented.

The reason it does not matter whether these calls were independent or not is simply that the method makes no assumptions regarding how they came to be distributed as they are. The analysis merely assumes that the order in which the five stimulus names are presented is random, and this requirement is fully met by the conditions. For example, the names might have been those of five

persons known to the group of subjects, and the wide variation in the number of names called might have reflected differences in popularity. The Greville test of significance would still give a correct probability figure showing how frequently, on a purely random basis, this highly biased set of responses on the five trials would correspond with the order of target names as well as was found to be true in the particular instance observed. If the probability figure meets the accepted criterion for statistical significance, the indication of a causal relation between the responses and the random order of target names is precisely as strong as in an experiment analyzed by any other method if the two sets of results happened to be significant at the same level.

Suppose that in this instance the random target order Brown, Hill, Smith, Greene, and Jones had been used. By summing the figures from the appropriate column for each trial, we find that the score of the group was 48. To determine whether or not this is significant, we only need to derive the mean and the variance by the Greville statistic, find the deviation of this score from the mean, and then divide the deviation by the standard deviation (square root of the variance) to arrive at a critical ratio for the test. All that this test of significance assumes is that the target order used was selected at random.

Keeping this illustration in mind, we may turn now to the question of how the Greville method of evaluation can be applied to verbal material obtained in token object tests. An example of its application to the records actually obtained in a set of five such tests will be given. For convenience, let us designate the cooperators as Jones, Brown, Smith, Hill, and Greene.

When the five records of this set were received from the subject, they were prepared for marking by the insertion of checking points within the verbatim statements. The five records were then arranged in random order and each one was given a code designation. They were then typed with sufficient copies and all five records were sent to each of the five cooperators to be checked throughout. The cooperators were given instructions to mark the items within each record on the assumption that they were all intended for them, and to use a check mark ($\surd$) to show a correct

statement, a cross (X) to show an incorrect one, and (?) for a doubtful item. In general, the instructions were so worded as to encourage a liberal interpretation rather than a restricted one. The aim was to get the widest possible measure of application of the material to the personal circumstances of each cooperator.

After the marked records were returned by all five of the cooperators, the items checked as correct were tabulated for each one, thereby showing the number of items he marked as correct in each record. The following distribution of check marks was found:

Records	Cooperators' Check Marks				
	Jones	Brown	Smith	Hill	Greene
First	9	7	0	0	3
Second	3	0	0	0	12
Third	10	0	26	3	13
Fourth	2	0	0	2	12
Fifth	3	0	0	4	4

These figures mean that Jones checked nine items as correct for his circumstances in the first record, three in the second, etc. These figures, reading them horizontally, could be thought of as meaning that the first record was called "Jones" nine times, "Brown" seven times, etc. Thus the figures take on the same significance as the distribution of subjects' calls in the illustration of the group test with five cards bearing the same names as shown on page 250 above. In fact, we have used identical figures in both illustrations to emphasize the close similarity of the two situations from a statistical point of view.

In the case of the token object test illustrated here, the "owners" of the five records were the targets. These were presented in one of the 120 possible permutations, an order that was selected at random and kept secret from the cooperators until after they had done their checking. It is this random order of "targets" that provides the basis for applying the Greville statistic, and the evaluation gives a straightforward statement of the probability value of the observed score obtained from the way in which the cooperators distributed their check marks.

In the example we have presented, the first record was Brown's, the second Hill's, the third Smith's, the fourth Greene's, and the fifth Jones's. The total score and the evaluation are exactly the

same as for the group card test with the random order of target cards assumed in that instance.

If significant results were obtained on verbal material appraised in this manner, the interpretation would be made in the same manner as for any test of significance. The method would permit the exclusion of chance as a reasonable explanation of the results, and this is all that any statistical analysis can do. If only five records were used in an analysis, it would not be possible to obtain a P-value of less than $\frac{1}{120}$ for a single set. When small sets are used, the results of a number of them combined might form the basis for any conclusion.

The reader who is interested in the more technical aspects of this method is referred to Greville's original article. In the appendix we have presented the basic formulas only as far as they are necessary for the evaluation of verbal material. Also, the complete evaluation of the data from the set of token object tests described in this section is presented in the appendix.

Suggestions

The following appear to us to be the most urgent research needs:

1. It is important to find out how sensitive the Greville method is when using different numbers of records in each analysis. As some of our colleagues have suggested, sets of five tests may be too small for the best results. But for convenience and speed in handling the results, small sets are advantageous. Therefore the aim should be to keep the sets as small as possible without making the test of significance too insensitive.

2. For the purpose of further refining the statistical practices, it is important to have material that will give significant results. Ideally, this should be obtained from actual experiments with psi capacities. But if successful subjects for token object tests are not available, verbal records might be deliberately made up to fit certain cooperators in order to see what size set is to be preferred. This

method has already been used with good effect in the Parapsychology Laboratory in comparing different methods of evaluation.

3. The psychological conditions, too, are important. The results obtained thus far suggest that in this respect the tests have not provided the essential requirements for success. For example, the use of sealed token objects may be an unfavorable way of working. It might be better to use exposed token objects that would not reveal anything regarding their owners, such as similar buttons or keys. Heretofore we have been sending all the token objects for a set of tests to a subject at one time, and there have been some indications that under these conditions a confusion among the cooperators comparable to the displacement effect may have occurred on several occasions. The results suggest that it might be preferable to send out only one token object at a time. Both patience and ingenuity are needed to devise test procedures that will satisfy the statistical requirements and provide favorable psychological conditions at the same time.

Appendix

The use of the Greville method in evaluating verbal material may be illustrated by the data from the set of token object tests given in the paper. We shall present the formulas in convenient computational form without going into technical questions of derivation and proof.

For the analysis, a slightly different arrangement of the figures from that shown on page 252 will be convenient. The columns showing the number of items the cooperators said were correct for themselves may be arranged across the page in the same order as the cooperators down the left-hand margin. The marginal totals of the figures within the matrix are needed, as well as the total number of times items were marked as correct. The data as prepared for evaluation are then as follows:

Cooperators	Cooperators' Check Marks					Total
	Brown	Hill	Smith	Greene	Jones	
Brown.....	7	0	0	3	9	19
Hill.....	0	0	0	12	3	15
Smith.....	0	3	26	13	10	52
Greene.....	0	2	0	12	2	16
Jones.....	0	4	0	4	3	11
Total.....	7	9	26	44	27	113

The number of items checked, or the sum of the figures in all 25 cells of the table, constitute the "calls." In this case, the number of calls is 113.

The number of "hits" is the sum of the figures on the main diagonal, or 48.

The mean number of hits expected is one-fifth of the number of calls, or 22.6.

This particular set thus gave 48 hits where 22.6 were expected, or a deviation of +25.4.

In order to compute the variance of the scores from the observed distribution, the following values are required:

The square of the number of calls (N^2) = 12,769

The sum of the squares of the individual cells (Σa^2) = 1,439

The sum of the squares of the separate row totals (Σr^2) = 3,667

The sum of the squares of the separate column totals (Σc^2) = 3,471

If n is the number of token object tests in a set, the variance is given by the expression:

$$V = \frac{1}{n^2(n-1)} [N^2 + n^2 (\Sigma a^2) - n(\Sigma r^2) - n(\Sigma c^2)]$$

For the present case we have:

$$\frac{1}{25(5-1)} [12{,}769 + 25\ (1{,}439) - 5\ (3{,}667) - 5\ (3{,}471)] = 130.54$$

The SD = $\sqrt{V}$ = 11.43

The CR of this set is therefore $\frac{+25.4}{11.34} = +2.22$

(P = .013)

When the results from a number of sets involving the same number of cooperators are combined, the mean number of hits for the total is the sum of the means of the individual sets, and the variance is the sum of the separate variances.

References

1. Greville, T. N. E. On multiple matching with one variable deck. *Ann. math. Statist.,* 1944, **15,** 432-34.
2. Pratt, J. G. *Towards a Method of Evaluating Mediumistic Material.* Bull. 23, Boston Soc. psych. Res., 1936.
3. Saltmarsh, H. F. Report on the investigation of some sittings with Mrs. Warren Elliott. *Proc. Soc. psych. Res.,* Lond., 1930, **39,** 47-184.
4. Saltmarsh, H. F., and Soal, S. G. A method for estimating the supernormal content of mediumistic communications. *Proc. Soc. psych. Res.,* Lond., 1930, **39,** 266-71.
5. Thomas, J. F. *Beyond Normal Cognition.* Boston: Boston Soc. psych. Res., 1937.

Parapsychology Laboratory
Duke University
Durham, North Carolina

Psychology Department
Duke University
Durham, North Carolina