ANALYZING FREE-RESPONSE DATA: A PROGRESS REPORT

JESSICA M. UTTS

Introduction

Free-response experiments are often preferable to forced-choice experiments, partly because they allow for the possibility of observing striking correspondences between the target and the response. Unfortunately, the statistical methods used to evaluate these experiments are not generally sensitive enough to allow full credit for such correspondences. These analysis methods are primarily adapted from methods used for forced-choice experiments. Thus, the degree of correspondence between the target and the response is often reduced to a very conservative approximation of true correspondence. One aim of this paper is to show that the same analysis ideas can be used in a new way to allow more credit for such correspondence.

Another problem with free-response experiments is that their complexity often leads to incorrect application of statistical methods (e.g., see Kennedy, 1979). This paper reviews some common analysis methods from the perspective of the assumptions necessary for their use. Also reviewed are sources of randomness that allow these assumptions to be met.

The paper is divided into two major sections. The first section shows how some common approaches to analyzing free-response data can be categorized according to the source of randomness built into the experiment or the analysis. The basic requirements given in the first section must be followed in order to apply these methods or extensions of them. That section also describes some pitfalls that must be avoided.

The second section presents some advances in the analysis of free-response (remote viewing) experiments at SRI International. These advances allow for more refined estimates of the degree of correspondence between the target and the response. As shown in this paper, these methods fit into the context of the sources of randomness dis-

cussed in the first section and can thus be viewed as extensions of the existing analysis techniques.

Sources of Randomness

It is well known that responses in psi experiments, both forced-choice and free-response, cannot be considered to be random in any sense. Yet, all statistical analyses are based on the assumption that the experiment or analysis contains some source of randomness. By categorizing methods for analyzing free-response experiments according to where the randomness enters the procedure, we may avoid incorrect analyses such as those discussed by Kennedy (1979). Focusing on the source of randomness may also help in the design of free-response experiments.

Past discussions of free-response analyses methods have distinguished between holistic and atomistic approaches (Burdick & Kelly, 1977, p. 110). In holistic approaches, a judge assigns rankings or ratings to responses matched with their corresponding targets and other potential targets. Atomistic approaches are those for which specific features are compared in the target and the response. As shown below, both types of analysis can be categorized by the nature of the underlying assumptions of randomness.

Sum-of-ranks method. A common procedure for analyzing free-response experiments is to ask a judge to assign a rank to each target/response pair, and then use the sum of ranks across trials as a summary measure. Stuart (1942), Morris (1972), and Solfvin, Kelly, and Burdick (1978) all discuss this method. Solfvin et al. list the assumptions needed for the application of this method as: "a) there is only one judge per trial; b) all targets are equally likely to be selected; and c) successive trials can properly be treated as independent" (p. 94).

For completeness and future reference, we present the formulas used to find significance levels for this approach. Let R = number of ranks possible for each trial, n = number of trials, and M = sum of ranks. Then the exact significance level for a given sum of ranks is

$$P = \frac{1}{R^n} \sum_{j=n}^{M} \sum_{k=0}^{\frac{j-n}{R}} (-1)^k \binom{n}{k} \binom{j - kR - 1}{n - 1}.$$

For large n, the sum is approximately normal with $\mu_M = n(R + 1)/2$ and $\sigma_M^2 = n(R^2 - 1)/12$. Thus, a z-score can be formed as $z = (M - \mu_M \pm .5)/\sigma_M$, and the significance level can be found in the normal table.

The assumptions (a through c) listed above are sufficient, but not

always necessary to ensure that these formulas are valid. The key to understanding when they apply is understanding the basic assumption built into the computation of the formulas (under the null hypothesis):

ASSUMPTION 1. *The summary statistic is a sum of integers. Each number in the sum is an integer from 1 to* R. *The* R^n *possible sets of integers that could make up the terms of the sum are all equally likely.*

In the situation where this technique is usually applied, the ranks are assigned on each trial by presenting a judge with the response, and with the correct target embedded with R-1 decoys. The key source of randomness in this application is that the target and each decoy must have been equally likely to have been the actual target at the start of the experiment. Furthermore, the rank assigned on any given trial must not influence or be influenced by the rank assigned on any other trial. It was this latter condition that was violated in some experiments discussed by Kennedy (1979); in those experiments the judges ranked each target against each response. Thus, assigning a target a rank of one for a particular response might have precluded that same target from being assigned a rank of one for another response.

Notice that it is crucial that the target and decoys were all equally likely to be chosen as the target at the beginning of the experiment, since the target selection was the only random source in the experiment. Thus, for example, if a target was selected without randomizing being involved and decoys were selected later, even if they were the same type as the target, the approach was invalid. In fact, since no randomization is involved in such an experiment, no statistical technique can be recommended without flaw.

We now describe another use for these formulas, which seems to have been unrecognized. R. G. Jahn, Dunne, & E. G. Jahn (1980) describe an atomistic approach to remote viewing analysis that uses a 30-bit descriptor list. They outline several normalization and scoring methods that can be used to access the quality of a particular remote viewing. The final analysis, however, is done by converting the quality measure to a rank. This rank is determined by computing the quality measure for a given response as matched against each possible target. If there are R possible targets, then R quality measures are computed for the given response. The rank assigned for the trial is simply the number of targets in the pool that match the response as well as or better than the actual target used in the trial. In other words, the target/response pair is assigned a rank, but it is assigned by a formula instead of by a human judge.

Jahn et al. then use what they call "the common z method for a

discrete distribution" (p. 223) to obtain significance levels. But this is simply the normal approximation given above for the sum-of-ranks method. (They apply this to the average rank instead of the sum of ranks.) Their summary statistic is exactly equivalent to a sum-of-ranks statistic. Given certain assumptions about how the experiment is conducted, the above formulas are valid. Thus, for example, in their Table 12 (p. 227) they use the normal approximation even though there are only five trials. The exact formula (1) given above could be used instead.

The crucial feature of the sum-of-ranks method, and thus the method used by Jahn et al., is that assumption 1 must hold. Under the null hypothesis, each rank must be equally likely to be any integer from 1 to R, independent of the ranks assigned during other trials. These conditions would hold for the following experiment, analyzed using an atomistic approach. A pool of N targets is selected and coded according to the bit list. A series of n trials is conducted by choosing targets from the pool with replacement. Each response is coded according to the bit list. Ranks are assigned by choosing a quality measure, computing it for the response compared to each of the N targets, and then counting how many targets match as well as or better than the actual target. (If ties are present, a slight modification is necessary. This is common if the quality measure can only assume a few values.) The sum of the ranks is computed, and the significance level is evaluated using formula (1) or the normal approximation.

This method is not valid if the targets are chosen without replacement, because assumption 1 no longer holds. To see this, suppose an experiment is conducted with only two targets, T_1 and T_2, in the pool and two trials. As an extreme case, suppose both responses yield the exact same bit-list configuration. Further, suppose this particular response configuration matches T_1 better than T_2, so that when T_1 is the correct target, a rank of 1 is assigned; and when T_2 is the correct target, a rank of 2 is assigned. If the targets are sampled without replacement, the only possible sets of ranks are (1,2) or (2,1). If sampled with replacement, all $R^n = 2^2 = 4$ possible sets are equally likely, and thus assumption 1 is met. Notice that the only source of randomness in the experiment is in the target selection; it is that source that allows or disallows the use of assumption 1.

Forced one-to-one matching. Forced-matching procedures for free-response experiments are discussed by Burdick and Kelly (1977), who attribute the first computation of the exact probability distribution to Chapman (1934). Scott (1972) gives tables that can be used to find significance levels; formulas for exact probabilities are given by Feller (1986, pp. 107–108).

The procedure is essentially equivalent to comparing two closed decks and counting the number of matches. For example, suppose N free-response trials are conducted and a judge is given the N targets and N responses, and told to match them one-to-one. The statistic of interest is the number of correct matches. There are $N!$ possible configurations of matches. The assumption used to calculate the formulas and tables mentioned above is

ASSUMPTION 2. *The summary statistic is the number of correct matches when matching* N *targets to* N *responses. Each of the* N! *possible configurations of matches is equally likely.*

This is generally the case (under the null hypothesis of no psi) if a closed set of N targets is presented in random order, and then the targets and responses are sufficiently randomized before being presented to the judge. Hyman and others suggest that problems can arise if trial-by-trial feedback is given (see, for example, Druckman & Swets, 1988, p. 182). Under such conditions, a subject might avoid mentioning features that were prominent in previous targets. Those targets would then have a smaller than average chance of being matched with the new response, thus negating assumption 2. Notice that while the source of randomness in this kind of experiment is the random presentation order of the targets, the other details of the experiment must ensure that assumption 2 holds under the null hypothesis of no psi.

Forced matching can also be employed in experiments using an atomistic, bit-list approach. A set of N targets can be coded according to the bit list, and then presented in random order over N trials. The quality measure derived by comparing the bit lists for targets and responses can be computed for all N^2 target/response pairings. Matching can then be done by finding the one-to-one pairing that maximizes the sum of the quality measure over the N pairs. The summary statistic is the number of correct matches in that pairing. This is essentially the method Scott (1972, pp. 86–87) recommends for evaluating verbal statements from mediums, except that the N^2 quality measures in that case are based on the N subjects' assessments of the accuracy of the statements in the N readings provided by the medium.

An interesting feature of the matching method is that the probability of exact x matches, and, thus, the significance level (significance level = P[x or more matches]), is about the same for any N of at least 10. Further, these probabilities are quite accurately approximated by the Poisson distribution with mean (and thus variance) of 1 (Feller, 1968, p. 108). The appropriate formula is

$$P(\text{exactly } x \text{ matches}) \approx e^{-1}/x! = .367879/x!,$$

regardless of N. Using this formula, P(4 or more matches) = .019, and P(3 or more matches) = .080. Thus, regardless of the number of trials, four or more matches lead to a significant result, while three or fewer do not!

Unforced matching. Many experiments criticized by Kennedy (1979) had the order of target presentation as the only source of randomness. Instead of being asked to do a one-to-one matching, however, judges were instructed to rank each response against all targets. The summary statistic and significance level were then based on either the sum-of-ranks method described above, or the number of hits as compared with a binomial distribution. Both methods assume trial-by-trial independence, a feature not present in these experiments.

The most conservative approach for reanalyzing these experiments correctly, and the one adopted by Kennedy, is to assume forced matching was used and then evaluate the significance level for the number of first-place matches. How conservative is this approach? It depends on the behavior actually adopted by the judge. The following discussion compares the two extremes when the summary measure is the number of direct hits.

Assume N targets are compared to N responses, and there are M first-place matches. At one extreme, assume forced matching was used. At the other extreme, assume ranks were assigned independently for each trial. Table 1 shows a comparison of results for these methods.

Notice that the p-values for the two methods get closer as N gets larger. It is an established fact that for large N and small p, the binomial distribution is well approximated by the Poisson distribution with mean Np (see Feller, 1968, p. 153). Thus, for large N the two methods are essentially equivalent.

Permutation methods. Permutation methods were apparently first ap-

TABLE 1

Comparison of Methods for Evaluating First-Place Matches

	Forced Matching	Independence
Mean, μ_M	1	1
Variance, σ_M^2	1	$(N-1)/N$
Distribution of M	Approx. Poisson	Binomial, $p = 1/N$
p-value, $N = 4$, $M = 4$	.042	.004
p-value, $N = 10$, $M = 4$	.019	.013
p-value, $N = 20$, $M = 4$	.019	.016

plied to free response data when Pratt and Birge (1948) recognized that Greville's (1944) forced-choice formulas were applicable to the assessment of verbal material from mediums. They restricted discussion, however, to methods using a normal approximation. Scott (1972, p. 87) seems to have been the first to recognize how to do an exact test.

Consider an experiment with n trials, so there are n targets and n responses. Suppose that judging is done by creating an $n \times n$ matrix of scores comparing each target with each response. These scores could be based on, for example, ranking the n targets for each response, using an atomistic quality measure for each target versus each response, or having a judge assign ratings to the degree of correspondence.

To apply the permutation method, arrange the matrix so that the scores for the correct matches are on the diagonal. The total score for the correct match is the sum of the diagonal elements (the trace) of the matrix. The summary statistic for the experiment is the proportion of all possible matches that have a total score as good as or better than the total score for the correct match. In other words, if the columns of the matrix are permuted in each of the $n!$ possible ways, and the trace of the matrix computed for each permutation, the summary statistic is the proportion of those traces that are as good as or better than the trace for the correct ordering. Note that in some cases, such as rankings, smaller traces are better; in other cases, such as ratings, larger traces are better.

To apply this procedure the order in which the targets are used must be randomized just as in the case of forced matching. The assumption under the null hypothesis can be summarized as follows:

> ASSUMPTION 3. *A series of* n *responses is given and compared to each of* n *targets to form an* n × n *matrix of scores. The summary measure is the proportion of permutations of targets for which the total sum of scores is as good as or better than for the correct ordering. At the start of the experiment* n! *possible orders of use of targets were equally likely.*

Notice that this technique is not appropriate if the order of use of targets is not random, although it may be tempting to try to use it in such a case. Non-psi factors such as the day's headlines and weather can be too easily incorporated into both the choice of the target and the response.

Remote Viewing Methodology

Humphrey, May, and Utts, (1988) discuss a methodology being developed at SRI International for the analysis of remote viewing ex-

periments. In this section, we first summarize the methodology, then show how it can be applied under the different assumptions in the previous section. Finally, we show how this methodology can be used to pick decoys for free-response experiments.

Quantitative definitions of targets and responses. The main goal in the analysis of remote viewing data is to assess how well the responses match their intended targets. To make that assessment, three elements are needed: a definition of the target, a definition of the response, and a measure of comparison.

Recent experiments in remote viewing at SRI have used an established pool of 200 photographs from *National Geographic*. Responses have been limited to a few pages of drawings and words. The purpose of the present analysis has been to develop a method of quantifying the targets and responses that is refined enough to incorporate both concrete and abstract features and that is flexible enough to allow the definition to be changed according to the purpose of the experiment, the level of experience of the subjects, and so on. In an experiment with novice subjects, for example, the goal might be to see if they can identify major features; in an experiment with more experienced subjects the goal might be to measure identification of more specific features.

To accomplish these goals, a list of 130 features was developed. These were categorized into ten levels, ranging from specific structures (e.g., churches, forts) in level ten, to abstract one-dimensional geometry (e.g., parallel lines, spirals) in level one. The complete list is given by Humphrey et al. (1988).

The 200 targets in the pool were coded according to the visual importance of each of the 130 features on the list. For each feature a value between 0 and 1 was assigned, with 1 meaning that the feature virtually dominated the entire picture, and 0 meaning that the feature was absent. Thus, the quantitative definition of a target consisted of a list of 130 numbers, each between 0 and 1, describing the degree of visual importance of each feature on the list.

After an experiment was conducted, the responses were coded similarly, except that the number assigned to each figure represented by the analyst's degree of belief that the feature was present in the response. For example, if the response contained the word *river*, then the river feature was assigned a value of 1. On the other hand, if the response contained a drawing of parallel snaking lines without a label, the analyst might have assigned a value of .3 to the river feature.

To compare the targets with the responses, the values assigned to the features should have the same meaning in both. Thus, for this

phase of the analysis, the target values were set to 1 for each feature for which the visual importance was rated at .2 or higher, since those features were definitely present in the target. The others were set to 0.

Comparison of targets and responses. May, Humphrey, and Mathews, (1985) describe a method of comparing targets and responses based on a *figure of merit* (FM). This measure is essentially a product of the proportion of the target material that was in the response (the *accuracy*) times the proportion of the target material that was correct (the *reliability*). The accuracy, reliability, and FM are easily adapted for comparing targets and responses as defined using the list of 130 features. The general versions of the formulas for the jth target/response pair are

$$\text{Accuracy}_j = a_j = \frac{\sum_k W_k(R_j \cap T_j)_k}{\sum_k W_k T_{j,k}},$$

$$\text{Reliability}_j = r_j = \frac{\sum_k W_k(R_j \cap T_j)_k}{\sum_k W_k R_{j,k}},$$

and $FM_j = a_j \times r_j$, where $R_{j,k}$ and $T_{j,k}$ are the values for feature k in response j and target j respectively, and $(R_j \cap T_j)_k$ is the intersection between the target and response for feature k, defined in this application to be $\min(R_{j,k}, T_{j,k})$. The sums are taken over all 130 features in the list. In this version of the figure-of-merit definition, we allow for the possibility of adding weights W_k, in order to change the contribution of various features to the FM.

Assessment of a single remote viewing. The quality of a single remote viewing can be assessed by computing FMs for the response compared to each of the 200 possible targets. Assuming that the target was elected randomly from the set of 200, with each target equally likely to be chosen, the proportion of FMs as large as or larger than the one for the correct target can be thought of as a p-value. It represents the probability—under the null hypothesis of no psi—of obtaining a match as good as or better than the one obtained.

Note that the crucial source of randomness here is the equal probability of selection for each target. Certain targets, particularly those with more detail, produce higher FMs on the average than others. The quality of a remote viewing, therefore, cannot be assessed by the magnitude of the FM alone.

Assessment of the entire experiment. An entire experiment based on n trials can be evaluated using one of the methods in the previous section, with the choice of method depending on how the experiment was conducted. Suppose a series of n trials is conducted and the targets are selected with replacement. The sum-of-ranks method can be used by computing the rank of the FM for the correct target when embedded in the ordered list of all 200 possible FMs. Under the null hypothesis, and assuming no ties, this rank is equally likely to be any integer from 1 to 200. To see this, suppose the response is generated before the target is selected. The 200 FMs can then be computed and put into an ordered list. The corresponding ranks from 1 to 200 can be assigned. Randomly selecting a target is equivalent to randomly selecting one of those ranks, with equal likelihood for each one. The argument does not change if the target is selected before the response is generated.

After conducting n such trials and finding the corresponding ranks, the significance of the sum of the ranks can be evaluated using equation (1), with $R = 200$, or the corresponding normal approximation. Note that the legitimacy of using the normal approximation is based on the magnitude of n, not R.

The other analysis methods discussed in the previous section can be used similarly. For example, if an experiment is conducted by selecting N targets from the pool of 200 and presenting them in randomized order without replacement, then the forced-matching method described for atomistic bit-list approaches can be used. Or, a matrix of FMs can be created and used in the permutation methods.

One problem with these approaches is that statistical power may be low because the FM depends on the target complexity. More complex targets are more likely to be matched to responses because of this dependency. This does not affect the significance level, but it may give unnecessarily discouraging results. Work is underway to try to normalize the FM to avoid this problem. Meanwhile, the feature list and cluster analysis have been used to help choose decoys for human judging.

Using the feature list to choose decoys. In addition to the problems already mentioned with the feature-list approach, certain elements in both the targets and the responses are not contained in the list of 130 features. Furthermore, no mechanism in the FM approach gives credit to responses that look similar to the target in various ways but are possibly mislabeled. So far, the best approach for evaluating such matches seems to be to use a human judge, presented with R-1 decoys embedded in a set with the correct target.

One issue of concern with the judging approach is how to choose

decoys that are dissimilar enough to not be confused with the correct target. For example, in the pool of *National Geographic* photographs there are several waterfalls, several snow-capped mountains, and so on. If decoys were selected from this pool randomly, there would be a relatively high probability that a decoy would look similar to the actual target. The following discussion presents a method of selecting decoys such that they are as dissimilar as possible.

The original assignments of visual importance of the 130 features can be used to compare targets and separate them into groups from which decoys can be selected. Similarity between targets can be assessed by computing an FM for the pair of targets. Using the same notation as in formula (2), we define the similarity ($S_{j,k}$) between targets j and k to be

$$S_{j,k} = \frac{\left(\sum_i W_i(T_j \cap T_k)_i\right)^2}{\sum_i W_i T_{j,i} \sum_i W_i T_{k,i}}.$$

Using these measures, we can create clusters of targets that are similar within clusters and different between clusters. For N targets there are $N(N-1)/2$ unique values (19,900 for $N = 200$) of $S_{j,k}$. The values j and k that correspond to the largest value of $S_{j,k}$ represent the two targets that look most similar. Suppose another target, m, is chosen and $S_{m,j}$ and $S_{m,k}$ are computed. If both values are larger than $S_{m,n}$ (for all n not equal to j or k), then target m is assessed to be most similar to the pair j, k. The process of grouping targets based on these similarities is called *cluster analysis*. See Johnson and Wichern (1982, Chapter 11) for a discussion of various clustering algorithms. We used hierarchial clustering with the complete linkage method, with the S-Plus statistical software package. Statistical packages such as BMDP and S also have clustering routines; BMDP has a version for PCs.

This procedure was followed to create clusters of the 200 targets in the *National Geographic* pool. Table 2 provides an overview of the 19 clusters found in the analysis. Some names appear to be quite similar, but, in fact, these sets are visually quite distinctive. Figure 1 shows the graphic output of a single cluster in detail. A much more complex—and visually difficult—graph is generated for the full cluster analysis and is not included here; this smaller subset has been chosen to illustrate the analysis. (To make the graphic analysis more meaningful, we did the analysis with $1 - S_{j,k}$.) All targets in this particular sample cluster are islands. Except for one outlier (i.e., a hexagonal building covering

TABLE 2

Names of the 19 Clusters

No.	Name	No.	Name
1	Flat Towns	11	Cities w/Prominent Geometries
2	Waterfalls	12	Snowy Mountains
3	Mountain Towns	13	Valleys with Rivers
4	Cities with Prominent Structure	14	Meandering Rivers
5	Cities on Water	15	Alpine Scenes
6	Desert/Water Interfaces	16	Outposts in Snowy Mountains
7	Deserts	17	Islands
8	Dry Ruins	18	Verdant Ruins
9	Towns on Water	19	Agricultural Scenes
10	Outposts on Water		

an island), the islands fall into two main groups: with and without man-made elements. The natural islands include three similar mountain islands, two sandbars, and two flat verdant islands.

Once these clusters have been created, decoys can be selected such that the R choices for judging, i.e. the target and the R-1 decoys, are each from separate clusters. This ensures that no decoy is too similar to the target or to another decoy. Since clusters have varying numbers of photographs, one should select R clusters with equal probability, and then select a photograph within each cluster.

Using cluster analysis to create target packets. The concepts of target similarity and cluster analysis can also be used to create sets of targets

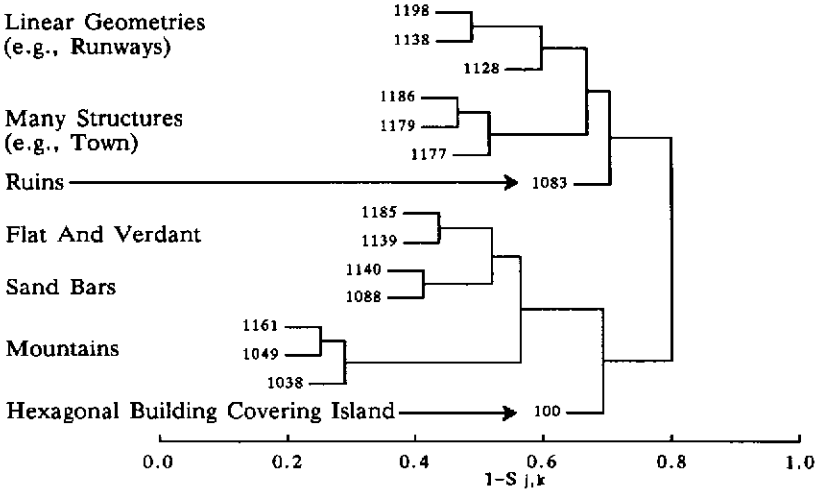

Figure 1. Detailed cluster analysis of the island cluster.

that are different within a set. Using this technique, we created 20 packets of 5 targets each from the *National Geographic* pool. To accomplish this, we used cluster analysis with *dissimilarity* between targets as the clustering criterion. Thus, the two most dissimilar targets were paired first, the next two most dissimilar next and so on, until a picture somewhat like Figure 1 emerged, but with all targets. Targets closest to each other were those most dissimilar. We used that information along with some visual shuffling to create packets of dissimilar targets.

These packets can be used as self-contained target/decoy units by randomly selecting a packet and then randomly selecting a target within the packet. Human judging can then be used by ranking the five targets in the packet against the response, repeating this for n trials with replacement and using the sum-of-ranks method of analysis.

REFERENCES

Burdick, D. S., & Kelly, E. F. (1977). Statistical methods in parapsychological research. In B. Wolman (Ed.), *Handbook of parapsychology* (pp. 81–130). New York: Van Nostrand Reinhold.

Chapman, D. (1934). The statistics of the method of correct matchings. *American Journal of Psychology, 45*, 287–298.

Druckman, D., & Swets, J. A. (Eds.). (1988). *Enhancing human performance*. Washington, DC: National Academy Press.

Feller, W. (1968). *An introduction to probability theory and its applications* (Vol.1, 3rd ed.). New York: Wiley.

Greville, T. N. E. (1944). On multiple matching with one variable deck. *Annals of Mathematical Statistics, 15*, 432–434.

Humphrey, B. S., May, E. M., & Utts, J. M. (1988). Fuzzy set technology in the analysis of remote viewing. *Proceedings of the Parapsychological Association 31st Annual Convention*, Montreal, Canada, pp. 378–394

Jahn, R. G., Dunne, B. J., & Jahn, E. G. (1980). Analytical judging procedure for remote perception experiments. *Journal of Parapsychology, 44*, 207–231.

Johnson, R. A. & Wichern, D. W. (1982). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice Hall.

Kennedy, J. E. (1979). Methodological problems in free-response ESP experiments. *Journal of the American Society for Psychical Research, 73*, 1–15.

May, E. C., Humphrey, B. S., & Mathews, C. (1985). A figure of merit analysis for free-response material. *Proceedings of the 28th Annual Convention of the Parapsychological Association*, Medford, MA, pp. 433–354.

Morris, R. L. (1972). An exact method for evaluating preferentially matched free-response material. *Journal of the American Society for Psychical Research, 66*, 401–407.

Pratt, J. G., & Birge, W. R. (1948). Appraising verbal test material in parapsychology. *Journal of Parapsychology, 12*, 236–256.

Scott, C. (1972). On the evaluation of verbal material in parapsychology: a discussion of Dr. Pratt's monograph. *Journal of the Society for Psychical Research, 46*, 79–90.

Solfvin, G. F., Kelly, E. F. & Burdick, D. S. (1978). Some new methods of analysis for preferential-ranking data. *Journal of the American Society for Psychical Research, 72*, 93–109.

Stuart, C. E. (1942). An ESP test with drawings. *Journal of Parapsychology, 6*, 20–43.

DISCUSSION

MORRIS: I have two kinds of comments. First of all I think it really is an interesting, innovative and exciting method. As far as the method itself is concerned, however, it sounds like an enormous amount of work and I wonder if you can comment on this aspect of it. What you have there, given all of the work that you put into it, is a very effective set of targets and sets of descriptors for them and there is now a lot that you can do with them. One question is: do you see this as something wherein there should be an attempt to have a standardized set of such targets that can then be used across labs? Do you see this as something wherein each lab that might define an idiosyncratic set of target materials really is going to have to do this all themselves right from scratch, including an enormous amount of playing around with the different levels of descriptors that they are going to try to deal with? If so, can you give some kind of feeling as to how easy this would be pragmatically for any lab to do given its own target interests? You have got geographical locations and sites and it seems to have worked really rather well with those. My second point is just simply more of a question. Do you see any restraints on the kinds of target pools that this might be used on just simply because it maybe harder to devise the different layers of meaning that you have been talking about?

UTTS: That is a good question. It certainly is a lot of work to put together a target pool and then to go ahead with the fuzzy set or with any bit list approach. Clearly it is a lot of effort. On the other hand, if a lab is in the situation where the experimenters know they are going to be in business for any length of time, which unfortunately is not always the case, then I think it is worth the effort. It is not valid to wait until the experiment is done and then fill out the bit list for the target, unfortunately. And so that is a problem. I think that you are right. It is a lot of work and it might be a useful idea to share target pools across labs. In fact that might add to the replicability issue. But you know it is like any endeavor; you have to decide how much the prep time is worth in the payoff at the end. I have to say frankly that I am not sure that the payoff for using the figure of merit approach was that high. I would say that there was strong payoff in using the method to choose decoys. When you have a huge target pool, such as 200 targets, it is hard to simply choose decoys and not get some repetition—you end up with a snowy mountain as the target and one of your decoys is also a snowy mountain and you are out of luck.

MORRIS: Can you share with us roughly how long it did take?

UTTS: I think that I will pass that on to Ed.

MAY: Well, it is really hard to say. We have been working out this problem for four or five years. I want to point out that it is really an iterative process. The pool started out at 400 and we did a clustering and then we noticed holes and some really junky targets that we could throw out. I would say with the target pool of 100 that we have now it would take maybe two years.

STANFORD: In terms of getting decoys that are "different" this seems great as a mechanical method. However, it seems to me that there may be a much more fundamental problem. This is derived from a set of categories that you developed from the bit list. But what does this really have to do with human cognition and perception? Similarly the way it works in the head is one thing and the way it works in terms of a system like this may be something altogether different. Do you have anything to say about that?

MAY: In fact, one of the real problems we had in putting this thing together was that Bev Humphrey, who was the primary mover on this for us, paid a great deal of attention to the particular bit list. Clearly the thing is sensitive as to what your bit list is. Under no circumstances would we recommend that this bit list be used in some other laboratory for two reasons. It was very highly tailored and I recommend that they all be tailored to match the target pool in question. Why have a purple giraffe in your bit list if there is not one in your targets. Also it was tailored to the general skill level of the subjects who were to be used in analyzing their data. That is a fair thing to do as long as you do it up front and a priori. So when you do that it is not such a large problem as you suggest. Also as to the fine-tuning of the bit, the target pool—which was frankly a surprise to me when we laid them all out on the floor—just looks visually different and that was our criteria, visual difference. And there was some minor fine-tuning on top of the technical part.

STANFORD: With regard to that, do you have ancillary evidence that that is true? Did you actually have people rating similarities so that you could show that?

MAY: Yes, we did. Certainly we did not do that over all 100, but we took samples of it and did it among ourselves around the laboratory. Our PA paper last year described in some detail how we gathered ground truth and did the comparison.

SCHOUTEN: I must say that I was very impressed by the paper. It is one of the things which has always interested me, because I think analyzing free-response data is really difficult. I have a couple of comments. One is that I think your approach to the bit list assumes that

the psi information would cover all of the details. A bit list means that you split up the target into details. If I understand you correctly, the scoring you use is, in effect, a combination: the more response items are correct, the more target items are correct, the higher the psi score. It is my impression from the analyses of spontaneous cases we carried out, that actually what often happens is that the basic concept, the idea behind an event is what is transmitted and not the details. But I must say it is an excellent way to establish differences between targets. A second comment is that I am a bit surprised because there are already various scaling methods which are used to establish distances between items. You might save yourself a lot of work using one of these if you only want to establish dimensions and different sets. A third point is with regard to sensitivity. In Utrecht we have been using a method which would give us a somewhat more correct test for evaluating free-response data. I grant immediately that yours are much better than ours. But then we took the data of an actual experiment and wrote a program to simulate outcomes of experiments in such a way that we introduced different levels of psi. So we increased the probability that the outcome was influenced by psi and then applied our evaluation method. We found, to our great disappointment, that the nice method that we had developed was still less sensitive in demonstrating the psi we had introduced into the data than the simple binomial. Did you ever try a simulation like that to find out whether your method is indeed more sensitive than the binomial?

UTTS: First, we did indeed find that the method of having human judges do rankings was more sensitive than using the figure of merit with the bit list approach as the actual assessment method. That is why we went back to just choosing the decoys using the fuzzy set approach. We used sum of ranks instead of a binomial, but it is the same idea to use that. Secondly, about other methods for making paired comparisons, we looked at some other methods and none of them seemed to do as well as this method. In fact, for that reason we were thinking of writing this method up and putting it out into literature in other fields where they are trying to solve the same problem. And finally, your first point was that just because you have done well by comparing things for this particular bit list does not mean that you have more psi. In response to that, you do have to tailor your list of things according to your definition of what you are looking for as evidence of psi. Now at SRI the decision was made that it should be visual correspondence. So this bit list is specifically designed to find visual correspondence.

HONORTON: In doing a meta-analysis of the ganzfeld work it was absolutely impossible to code anything concerning targets and com-

position of target pools. Just as it is theoretically possible, that a lot of the experimenter effect is due to different subject populations, it is quite possible that a lot of the variability in free-response studies is due to differences in free-response target pools. There is also the degree to which different investigators are successful in creating interesting and yet relatively orthogonal target pools. The situation is, I think, analogous to what would happen in psychology if, in Rosenthal's person perception test of experimenter expectancy effects, every investigator used a different set of photographs without describing anything about their characteristics or if clinicians using projective techniques such as the Thematic Apperception Test or the Rorschach Test each created his own ink blots or ambiguous figures. That is bound to greatly increase the variability. Certainly some degree of standardization is really very important.

PALMER: It seems to me that the more holistic methods, such as the matching procedures, and the more atomistic methods such as the one that you have developed are tapping very different things. Have you looked at the correlation between the results with those two methods? If the correlation is high, what would you think about possibly combining the two outcomes to get something that takes advantage of what is going on with both procedures?

UTTS: That is a good question. I would say the correlation is not that high because they are tapping different parts of what is going on. In fact, we have been looking at methods of trying to combine them and have not yet come up with one that we feel is satisfactory. But that is what we actually are ultimately trying to do.

MAY: One of the problems with this kind of atomistic approach and the holistic methods and rank order procedure is that if the response is of bad quality, but good enough that a judge can just squeak it into a first place match that is one circumstance. The second circumstance would be if you have a fantastically high correspondence, an agreement with the first place, and the judge had no trouble making a first place match, the statistic does not differentiate between that really great hit and that just squeaking-by hit. I feel that it is not fair to a good response, so we are looking at a way of merging the two procedures to take advantage of that.

STANFORD: When you talk about standardization as Chuck was—and I agree with those remarks—before we standardize let us be sure we have all the elements together. For instance, if I were myself going out after targets, I don't know whether I would get into some of these 19 that are listed. My own experience suggests—and this is a very clinical type of thing—that you can't frame scientifically, but that there

are some kinds of targets that might be a lot better than this. Some investigators, Chuck in his lab and some others historically, have been looking at what types of material make better targets. So when we standardize, if our aim is really high psi yield and not some specific kind of target that we are interested in or something of that sort, it seems to me that we really do need a lot more research on what types of targets individuals are likely to be sensitive to.

UTTS: I absolutely agree with you.

HONORTON: I also have been thinking a lot lately that what we maybe ought to do, at least in our experiments with novices, with people who have not done these free-response procedures before, is to have a single target pool that is used consistently across all of the screening or first-timers' sessions. One of the real problems in doing any kind of process-oriented research with free-response methods is that, to the extent that there are target effects and given the amount of time it takes to do an individual's free-response session, you can very well mask either good subjects or some correlate of subject performance, and because of the luck of the draw you get a psi-missing target. But if you have a standardized single pool that is used consistently for at least the initial stage, then all subjects are being assessed on an equal footing.

CARPENTER: I am new myself at free-response work, so my comment may be a bit naive, but I am wondering about limitation of the "bit" approach to analysis. What we have been doing is using group psychotherapy sessions as the mode of ESP response and then relating that to one of four pictures taken from magazines. The relationship between those two things reminds me of the dream work in that the relationships are very allegorical and metaphorical. For example, a session might have a certain mood and there might be no literal reference to anything that happened in the session in any element in the picture, but there is something about the mood that members of the group take as alluding to a kind of similarity. Now it seems to me that a bit approach would not have any way of catching that. I am wondering if those of you who have been doing this feel that anything significant is lost with the bit approach, the more metaphorical kind of relationship.

UTTS: I would say that you need to construct the bit list if you can to somehow incorporate those elements that you think are likely to arise that show evidence of psi. In the SRI case we were mainly focusing on visual correspondence. I do not know if there is a way to capture that sort of thing in a bit list, but that is something that you would want to think about. Then you need an experienced judge who can pick those out of the response.

MAY: I am not so sure that standardization of targets is the great

way to go, other than as Chuck just suggested, for the very first level novice activity. What is important with this particular procedure, at least from our point of view, is that one can tailor it to match whatever one is looking for. If you are really interested, Jim, in the allegorical nature of what you are doing, you can design a bit list that focuses upon that and down-plays the visual or maybe the literal interpretation, so you can explore that. In fact, the method is powerful enough to allow you to put different weighing factors in so you can explore specific imagery. The neat thing about this procedure, at least from my point of view, is that it is infinitely flexible and each group can tailor it to their own specific needs.

HONORTON: In fact, with a computer you get rid of visual targets altogether and tailor your bit descriptive list to the particular subject population, the kind of problem that you are working with. You can give the agent whatever the bit categories are that are selected for the session and let him or her create out of a playroom full of materials some representation of that. That is another possibility that would greatly increase the freedom of expression while still providing an objective scoring structure.