# MULTINOMIAL RUNS TESTS TO DETECT CLUSTERING IN CONSTRAINED FREE RECALL

Gail Rubin[1], Charles E. McCulloch[1]
and Michael A. Shapiro[2*]

Biometrics Unit[1], and Department of Communication[2], Cornell University, Ithaca, NY 14853

# ABSTRACT

Psychologists often want to detect category structure in subjects' free recall protocols. While runs tests based on the binomial distribution are commonly used to detect non-randomness within a sequence, many research situations require tests based on the multinomial distribution. We propose a test of randomness versus clustering based on the number of runs in multinomial data. We illustrate its use with data from a mass communication experiment using a constrained free recall procedure.

*Gail Rubin is Visiting Fellow, Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, NY 14853. Charles E. McCulloch is Associate Professor of Biological Statistics, Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, NY 14853. Michael A. Shapiro is Assistant Professor, Department of Communication, 640 Stewart Avenue, Cornell University, Ithaca, NY 14850-3899. This paper is Technical Report #BU-976-M in the Biometric Unit Series. The authors are grateful to D. S. Robson, C. Clogg and the reviewers of this paper for their helpful comments.

## 1. INTRODUCTION

To draw inferences about mental processes, psychologists often want to detect category structure in free recall. Typically, a list of randomly ordered words is presented to a subject. Unbeknownst to the subject, the experimenter has selected the words from a limited number of categories. At some later time the experimenter asks the subject to write down as many of the words as he or she can remember—a free recall procedure. The experimenter then looks for evidence that the subject used the underlying categories in recalling the words. Typically some measure of clustering by category is used (Murphy and Puff, 1982).

The study list does not have to be an artificial list drawn up specifically for an experiment. The investigator can ask the subject to call on real-life experiences as well. For example Shapiro (1987; 1988) was interested in whether subjects used mass media categories in classifying memories. To test that, he asked subjects to free recall specific people from certain categories typically over or under represented in the mass media (e.g., criminals or old people). If a subject used communication source categories in retrieving that kind of information, then the free recall items should cluster according to communication source (i.e., television fiction, a newspaper, a book). A multinomial runs test seems natural for detecting clustering in the constrained free recall setting, where no presented list exists.

In Shapiro's (1987; 1988) experiment, each of 155 subjects was asked to list all specific examples s/he could recall of a given kind of person in a fixed time period. This was done for each of 6 categories (topics): criminals, law enforcement personnel, victims of crime, professionals, 25-45 year-olds and old people. After the 6 lists were made, each subject was asked to classify each item on each list as to one of 8 possible sources from which information was obtained: direct experience, other people, books, newspapers, television news, television fiction, other television and movies. All 155 subjects were tested independently, the order of topics for each subject was varied using latin squares, and each subject made separate lists for each topic.

Hence, each list represents a sequence of memories, which can be tested for clustering of items according to communication source. Using a one-tailed multinomial runs test, one rejects (fails to accept) the null hypothesis of randomness of recall on a topic with respect to communication source, if the number of runs is too small. The multinomial runs test has the advantage of allowing each subject–topic combination to have its own vector of probabilities for the 8 sources. In Section 4, we will apply the results of Sections 2 and 3 to data from Shapiro's (1987; 1988) experiment.

Shapiro's constrained free recall experiment was designed to obtain indirect evidence for (or against) a proposed model of how information from various communication sources is stored and used. The model assumes that information about the communication source is stored along with an event memory. For example, the model claims that when a viewer stores a memory of a criminal from a television detective show, the communication source of that memory (television fiction) is associated with that memory. The spreading activation theory of memory (Collins and Loftus, 1975) predicts that remembering an exemplar from one source would activate other memories connected with that source. That activation would make it more likely that memories from that source would be recalled next. According to this model, in constrained free recall, consecutive items from the same communication source are more likely than one would expect by chance. Thus, one would expect fewer runs in the recall data than expected by chance. Observing many runs in the experiment would indicate that the model described above is inappropriate, while observing a statistically significant paucity of runs only would indicate that the model is a possible (but not unique) explanation of the mental processes.

To ascertain whether a relationship exists between an item's recall position and some other factor in free recall protocols, Pellegrino and Hubert (1982) suggest a statistic of the form

$$\Gamma = \sum_{i=1}^{n} f(x_i, y_i) \quad ,$$

where $y_i$ designates the position of the $i^{th}$ item in the recall list, $x_i$ designates some other characteristic of the $i^{th}$ item and $f(x_i, y_i)$ is a function specified by the experimenter. Pellegrino and Hubert (1982) focus on normalized correlation coefficients, which incorporate comparison of the recalled list with the presented list; they do not suggest a specific function useful in measuring clustering. Robertson (1985) suggests using the multinomial runs distribution to quantify clustering in free recall experiments by reporting the probability, under the null hypothesis of no clustering, of obtaining the same or fewer runs than that observed in the recall list. Notice that with data from the constrained free recall experiment correlation coefficients cannot be used, since no presented list exists, and a runs statistic cannot be written in the form suggested by Pellegrino and Hubert. We examined the utility of a test based on the multinomial runs distribution as a test of randomness versus clustering in such situations.

The distribution theory of runs developed out of interest in testing whether a sequence of events exhibits randomness rather than clustering of like elements within the sequence. According to Mood (1940), the early work began in the late nineteenth century with Karl Pearson and was motivated by interest in games of chance. Mood's 1940 paper gives the distribution theory for random arrangements of a fixed number of each kind of element as well as for random arrangements of a random number of each kind of element (i.e., distributions of elements from binomial and multinomial populations). Even today, runs tests are the primary tools available to distinguish clustering of like elements from randomness within a sequence (Lehmann, 1975; 1983), although the multinomial version of the runs test is not commonly discussed in statistical methods books and tables of critical values are not commonly available.

Shaughnessy (1981) gives the recursion formulae needed to calculate the exact distribution of runs and tabulates the critical values for the multinomial runs test with few classes (2-6) and approximately equal numbers in each class. Schwager (1983) gives formulae for calculating run probabilities for the more general case of sequences of Markov-dependent trials. Robertson (1985) and Koppen and Verhelst (1986) discuss the behavior of large sample approximations to the exact distribution of runs; both papers focus on fixed numbers of each kind of element, occurring in a free recall context. With even a moderate number of classes, calculation of the exact distribution becomes prohibitive.

To use Mood's results as the basis of runs tests, one must assume that the probability of the $i^{th}$ class ($p_i$) is known. In a binomial runs test it is often sensible to take $p_i = \frac{1}{2}$; however, in the multinomial situation, the $p_i$'s are usually unknown and must be estimated.

In Sections 2 and 3, we investigate the effect that estimation of the $p_i$'s has on the performance of the multinomial runs test. Mood, Graybill and Boes (1974, pp. 519-521) suggest a binomial runs test, which uses the mean and variance of the number of runs that are conditional on the observed number of elements of each type. A similar test could be devised for the multinomial case. Such tests are expected to be conservative. However, deriving estimators based on the conditional distributions would be unwise for the constrained free recall application. One would be leery of drawing inference, conditional on the observed n vector, when the given n vector is of no interest and unlikely to be observed again, even if one retested the same subject.

## 2. MULTINOMIAL RUNS TEST STATISTIC

In applying Mood's results to our problem, two major modifications were necessary. First, the asymptotic mean depends on the unknown $p_i$'s and therefore must be estimated. This estimation changes the asymptotic distribution. Second, the asymptotic results given in Mood (1940) were inaccurate for the sample sizes we encountered in practice; thus the exact moments were needed.

The second problem is somewhat simpler, so we deal with it first. Mood [1940, equation (7.20)] lists the means, variances and covariances of the $r_i$'s, the runs of each type, but does not explicitly list the mean or variance of the total number of runs, $r = \sum r_i$. The mean can be easily calculated as

$$\begin{aligned} \mathrm{E}[\mathrm{r}] = \mu_\mathrm{r} &= \textstyle\sum \mathrm{E}[\mathrm{r_i}] = \sum\Big[\mathrm{np_i(1-p_i) + p_i^2}\Big] \\ &= \mathrm{n\ (1-\textstyle\sum p_i^2) + \sum p_i^2}\ , \end{aligned} \tag{2.1}$$

where $\mathrm{p_i}$ is the probability associated with the $\mathrm{i}^{th}$ kind of element. The asymptotic result of $\mathrm{n(1 - \sum p_i^2)}$ can easily be incorrect by more than 10% of the standard deviation of r for small to moderate sample sizes.

The variance of r is given by

$$\begin{aligned} \mathrm{Var(r)} &= \sigma_\mathrm{r}^2 \\ &= \sum_\mathrm{i} \mathrm{Var(r_i)} + \sum_{\mathrm{k} \neq l}\sum \mathrm{Cov\ (r_k, r_{\mathit{l}})} \\ &= \sum_\mathrm{i}\Big[\mathrm{np_i\ (1 - 4p_i + 6p_i^2 - 3p_i^3) + p_i^2(3-8p_i - 5p_i^2)}\Big] \\ &\qquad + \sum_{\mathrm{k} \neq l}\sum\Big\{\mathrm{-np_k p_{\mathit{l}}(1 - 2p_k - 2p_{\mathit{l}} + 3p_k p_{\mathit{l}}) - p_k p_{\mathit{l}}(2p_k + 2p_{\mathit{l}}\ - 5p_k p_{\mathit{l}})}\Big\} \\ &= \mathrm{n}\Big\{\textstyle\sum \mathrm{p_i^2} + 2\sum \mathrm{p_i^3} - 3(\sum \mathrm{p_i^2})^2\Big\} + \Big\{-\sum \mathrm{p_i^2} - 4\sum \mathrm{p_i^3} + 5(\sum \mathrm{p_i^2})^2\Big\}\ . \end{aligned} \tag{2.2}$$

Again, leaving out the second term in braces to get the asymptotic result often causes an error of more than 10% of the standard deviation. This extra piece therefore must be included for the sample sizes we encountered.

The more serious complication is the need to estimate the unknown $\mathrm{p_i}$'s. Mood (1940) gives the result

$$\frac{\mathrm{r}-\mu_\mathrm{r}}{\sigma_\mathrm{r}} \sim \mathrm{AN(0,1)}\ ,$$

where $\mu_\mathrm{r}$ is given by the order n term in (2.1) and $\sigma_\mathrm{r}^2$ is given by the order n term in (2.2). In attempting to use $\frac{\mathrm{r}-\mu_\mathrm{r}}{\sigma_\mathrm{r}}$ to test for randomness, the $\mathrm{p_i}$'s are nuisance parameters and must be estimated. If $\hat{\sigma}_\mathrm{r}$ denotes a consistent estimate of $\sigma_\mathrm{r}$ (we will use the MLE) and $\hat{\mu}_\mathrm{r}$ denotes an estimate of $\mu_\mathrm{r}$, then it is well known that the asymptotic distributions of $\frac{\mathrm{r}-\hat{\mu}_\mathrm{r}}{\sigma_\mathrm{r}}$ and $\frac{\mathrm{r}-\hat{\mu}_\mathrm{r}}{\hat{\sigma}_\mathrm{r}}$ are equal. However, the distributions of $\frac{\mathrm{r}-\hat{\mu}_\mathrm{r}}{\sigma_\mathrm{r}}$ and $\frac{\mathrm{r}-\mu_\mathrm{r}}{\sigma_\mathrm{r}}$ are not asymptotically equal.

In particular

$$\begin{aligned} \mathrm{Var}\Big(\frac{\mathrm{r}-\hat{\mu}_\mathrm{r}}{\sigma_\mathrm{r}}\Big) &= \mathrm{Var}\Big(\frac{\mathrm{r}-\mu_\mathrm{r}}{\sigma_\mathrm{r}}\Big) + \mathrm{Var}\Big(\frac{\mu_\mathrm{r}-\hat{\mu}_\mathrm{r}}{\sigma_\mathrm{r}}\Big) + 2\mathrm{Cov}\Big(\frac{\mathrm{r}-\mu_\mathrm{r}}{\sigma_\mathrm{r}}\ , \frac{\mu_\mathrm{r}-\hat{\mu}_\mathrm{r}}{\sigma_\mathrm{r}}\Big) \\ &= 1 + [\mathrm{Var}(\hat{\mu}_\mathrm{r}) - 2\mathrm{Cov}(\mathrm{r},\hat{\mu}_\mathrm{r})]/\sigma_\mathrm{r}^2\ . \end{aligned} \tag{2.3}$$

The term in brackets will usually be nonzero.

To construct a test statistic, we propose replacing the value of $\mu_r$ by its unbiased estimate, $\hat{\mu}_r$, and calculating the asymptotic variance of r $-\hat{\mu}_r$. Explicitly, we suggest the use of

$$T= \frac{r-\hat{\mu}_r}{\sqrt{\hat{Var}(r-\hat{\mu}_r)}}\ , \tag{2.4}$$

where $\hat{\mu}_r = n(1-\sum \hat{p}_i^2) + 1$, which is unbiased for (2.1). It remains to calculate $Var(r-\hat{\mu}_r)$. From (2.3) we need to calculate $Var(\hat{\mu}_r)$ and $Cov(r,\hat{\mu}_r)$. The variance term is given by

$$\begin{aligned} Var(\hat{\mu}_r) &= Var[n(1 - \textstyle\sum \hat{p}_i^2)] \\ &= \frac{1}{n^2}\Big(\sum_i Var(n_i^2) + \sum_{k\neq l}\sum Cov(n_k^2, n_l^2)\Big)\ , \end{aligned}$$

where $n_i$ is the observed number of elements of the $i^{th}$ kind, and the variances and covariances in this last expression are given in Johnson and Kotz (1969, p. 284). The covariance term is slightly more tedious to calculate:

$$\begin{aligned} Cov(r,\hat{\mu}_r) &= Cov[r, n(1 - \textstyle\sum \hat{p}_i^2)] \\ &= -\frac{1}{n}\sum_i Cov(r, n_i^2) \\ &= -\frac{1}{n}\Big(\sum_i E[rn_i^2] - E[r]E[n_i^2]\Big)\ . \end{aligned}$$

In this last formula, E[r] is given by (2.1) and $E[rn_i^2]$ can be found as

$$\begin{aligned} E[rn_i^2] &= E\left[E[rn_i^2|n]\right] \\ &= E\left[n_i^2E[r|n]\right] \\ &= E\left[n_i^2 \sum_j \frac{n_j(n-n_j+ 1)}{n}\right] \text{ from (Mood, 1940)} \\ &= \frac{1}{n}\Big((n^2 + n)E[n_i^2] - E[n_i^4] - \sum_{j\neq i} E[n_i^2n_j^2]\Big)\ . \end{aligned}$$

Combining the calculations, we have

$$\begin{aligned} Var(\frac{r-\hat{\mu}_r}{\sigma_r}) = 1 + \Big\{ 2\, [(1-\tfrac{1}{n})(\textstyle\sum p_i^2) + \tfrac{1}{n}]\, [\sum_i E(n_i^2)] \ + \frac{1}{n^2}\Big( - \sum_i E[n_i^4] \\ - \sum_i E[n_i^2]^2 + \sum_{k\neq l}\sum E[n_k^2n_l^2] - 3 \sum_{k\neq l}\sum E[n_k^2]\ E[n_l^2]\Big)\Big\} \ /\ \sigma_r^2\ . \end{aligned}$$

This can be calculated in a straightforward, though tedious manner for use in practice.

Exact formulae are given in Appendix 1.

## 3. PERFORMANCE OF THE TEST STATISTIC

We investigated the performance of the test statistic given by (2.4) by simulation. We were especially concerned about its mean and variance under the null hypothesis, since in Section 4 we combine the number of runs calculated from separate subjects to form an overall test statistic. Details of the simulation techniques are given in Appendix 2.

The results of the simulation were encouraging. Even for samples of size n = 25 with 8 categories the standard deviation was close to one and the distribution was well approximated by a standard normal. For small sample sizes the only failing was a light tail on the right, leading to a slight conservatism for small $\alpha$ levels.

For the simulations, we considered probability vectors that were motivated by the constrained free recall experiment and spanned a wide range of values. The vectors considered were

$$
\begin{aligned}
p_1 &= (\,.125, & .125, && .125, && .125, && .125, && .125, && .125, && .125\,)' \\
p_2 &= (\,.411, & .204, && .112, && .106, && .095, && .032, && .024, && .016\,)' \\
p_3 &= (\,.586, & .170, && .111, && .057, && .038, && .017, && .013, && .008\,)' \\
p_4 &= (\,.443, & .306, && .107, && .070, && .030, && .024, && .014, && .006\,)' \\
p_5 &= (\,.289, & .280, && .148, && .124, && .050, && .048, && .039, && .022\,)' \\
p_6 &= (\,.441, & .184, && .149, && .078, && .065, && .041, && .026, && .016\,)' \\
p_7 &= (\,.182, & .172, && .169, && .121, && .112, && .104, && .074, && .066\,)'\,.
\end{aligned}
$$

In Table 1, we present the standard deviations of $T = \frac{r-\hat{\mu}_r}{SE(r-\hat{\mu}_r)}$ and $Z = \frac{r-\hat{\mu}_r}{\hat{\sigma}_r}$ for samples of size n = 10, 25, 50 and 100 for the seven probability vectors given above.

The results show that T is always an improvement over Z. The effect of estimation of $\mu_r$ is to greatly reduce the standard deviation from the nominal value of one. This would lead to extremely conservative tests. Often T is a substantial improvement over Z, e.g.,

$n = 50$ for $p_3$. Even for sample sizes as small as 10 or 25, T had a standard deviation close to one.

Table 1: Simulated standard deviations of $\frac{r-\hat{\mu}_r}{SE(r-\hat{\mu}_r)}$ and $\frac{r-\hat{\mu}_r}{\hat{\sigma}_r}$ for various sample sizes and probability vectors.

Standard Deviations

$$\left(\frac{r-\hat{\mu}_r}{SE(r-\hat{\mu}_r)}, \frac{r-\hat{\mu}_r}{\hat{\sigma}_r}\right)$$

Probability Vectors

| Sample Size | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ |
|---|---|---|---|---|---|---|---|
| 10 | .85,.70 | .91,.69 | .91,.61 | .96,.73 | .91,.72 | .93,.69 | .88,.72 |
| 25 | .92,.83 | .96,.77 | .89,.60 | .95,.76 | .94,.80 | .93,.72 | .92,.83 |
| 50 | .96,.91 | .99,.78 | .94,.60 | .97,.78 | .97,.84 | .94,.72 | .93,.86 |
| 100 | .96,.93 | .99,.78 | 1.00,.62 | .99,.80 | 1.02,.90 | .99,.76 | 1.01,.96 |

In Table 2, we present some selected tail probabilities for $p_6$, which had neither the best nor worst approximation to a standard normal distribution. The left tail probabilities are almost exactly equal to nominal, while the right tail probabilities are slightly conservative for small sample sizes.

Table 2: Simulated tail probabilities of $\frac{r-\hat{\mu}_r}{SE(r-\hat{\mu}_r)}$ for $p_6$ .

| | | Tail Probabilities | | | | | |
|---|---|---|---|---|---|---|---|
| | | Left | | | Right | | |
| Sample Size | Nominal | .010 | .025 | .050 | .050 | .025 | .010 |
| 10 | | .010 | .029 | .059 | .015 | .010 | .001 |
| 25 | | .008 | .023 | .047 | .031 | .011 | .003 |
| 50 | | .009 | .021 | .047 | .035 | .011 | .003 |
| 100 | | .012 | .025 | .053 | .051 | .021 | .010 |
| Simulation Standard Errors | | (.0026) | (.0040) | (.0056) | (.0056) | (.0040) | (.0026) |

## 4. APPLICATION TO THE CONSTRAINED FREE RECALL EXPERIMENT

In the constrained free recall experiment (Shapiro, 1987; 1988), each of 155 subjects was asked to list all specific examples s/he could recall on a given topic (e.g., criminals) in a fixed period of time. After a list was made for each of 6 topics, each subject was asked to classify each item on each list as to one of 8 possible communication sources. Hence, every list represents a sequence of memories, which can be tested for clustering of items according to information source. The multinomial runs test has the advantage of allowing each subject-topic combination to have its own vector of probabilities for the 8 sources. This is critical, since the likelihood of obtaining information from a given source will differ from topic to topic and is likely to differ from person to person. A global test can be performed by summing the number of runs over people and topics. Alternatively, one can combine over topics to test for randomness in sequences for the subjects or combine over subjects to test for randomness in sequences for the topics.

A statistic, T, that combines topics or combines subjects is calculated as follows:

$$T= \frac{\sum_{j}(r_{.j} - \hat{\mu}_{rj})}{\sqrt{\sum_{j}\hat{\text{Var}}(r_{.j} - \hat{\mu}_{rj})}}$$

where j indexes the categories to be combined, $r_{.j} = \sum_{i} r_{ij}$ and $\hat{\mu}_{rj} = \sum_{j}[n_j(1 - \sum_{i}\hat{p}_{ij}^2) + 1]$ . A similar statistic, based on Z, can be formed. A one-tailed test is appropriate for this problem: if there is clustering of items from different sources, there will be too few runs.

Table 3 illustrates the calculation of the runs statistic based on Mood's results, Z, and our runs statistic, T, using the data for one subject on two topics. Z is conservative for both topics. For criminals, one would reject the hypothesis of randomness of recall with respect to communication source based on either statistic. However, one draws different conclusions from the two statistics for law enforcement personnel. Of the 18 items for this topic, 14 were attributed to TV fiction and they occurred in 2 runs; the total number of runs for this topic was 5. The T statistic rejects the randomness hypothesis, while the Z statistic is too conservative, failing to reject the randomness hypothesis although the total number of runs is small.

The constrained free recall experiment also can be used to illustrate a runs test, which combines over subjects for a given topic. A simple random sample of subjects, with sample size 28, was drawn from the data base. Combined tests based on the T and Z statistics were calculated for the topic criminals. Using the data from the 28 subjects, $T_C = -12.48$ ($p = 0.0000$) and $Z_C = -9.82$ ($p = 0.0000$). One rejects the hypothesis of randomness of recall with respect to communication source based on either statistic. Again, the Z statistic is too conservative.

Table 3: Calculation of Z and T statistics for one subject and two topics from the Shapiro constrained free recall experiment. ($n_i$ = number of items from each source, $r_i$ = number of runs of each kind)

| | | Criminals | | | Law Enforcement Personnel | | |
|---|---|---|---|---|---|---|---|
| i | Source | $n_i$ | $r_i$ | $\hat{P}_i$ | $n_i$ | $r_i$ | $\hat{P}_i$ |
| 1 | TV fiction | 5 | 2 | 5/19 | 14 | 2 | 14/18 |
| 2 | TV news | 6 | 3 | 6/19 | 0 | 0 | 0 |
| 3 | other TV | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | movies | 2 | 1 | 2/19 | 0 | 0 | 0 |
| 5 | books | 1 | 1 | 1/19 | 0 | 0 | 0 |
| 6 | newspapers | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | direct experience | 1 | 1 | 1/19 | 3 | 2 | 3/18 |
| 8 | other people | 4 | 2 | 4/19 | 1 | 1 | 1/18 |
| Total | | n = 19 | r = 10 | | n = 18 | r = 5 | |

$Z = \frac{r-\hat{\mu}_r}{\hat{\sigma}_r} = -3.033$ | $Z = -1.0255$

p–value of Z = 0.0012 | p–value of Z = 0.1526

$T = \frac{r-\hat{\mu}_r}{SE(r-\hat{\mu}_r)} = -3.4799$ | $T = -2.1023$

p–value of T = 0.00025 | p–value of T = 0.0221

## 5. DISCUSSION

Based on the data from the random sample of 28 of the 155 subjects tested in the constrained free recall experiment, we conclude that there is a statistically significant paucity of multinomial runs in the recall lists. Thus, the null hypothesis of randomness of recall with respect to communication source is rejected, and the clustering of recall items according to information source is deemed significant. The six tests for topics, which combine over subjects to test for randomness in sequences for each of the topics, all indicate significant clustering of recall items according to communication source, at $\alpha = 0.01$. Under the null hypothesis of randomness, it may be reasonable to assume that topics, as well as subjects, represent independent data and a global test, calculated by summing the number of runs over subjects and topics, can be formed. Such a test for the six topics and 28 subjects easily rejects the hypothesis of randomness ($p = 0.0000$). If the assumption of independence of topics is untenable, a global test can be performed using Bonferroni corrections.

The significant clustering of recall items with respect to communication source in Shapiro's data is consistent with the psychological model, based on the spreading activation theory of memory, described in Section 1. However, other interpretations of the clustering are possible. The model in Section 1 suggests that the source of a memory activates other memories connected to that source. This process could occur in two different ways. During the initial phase when items are recalled, subjects might use the communication source as a recall strategy, generating examples from the same source in sequence. Alternatively, during the labeling phase when sources are ascribed to items in the recall list, subjects might tend to think of the same context (communication source) for each example. That is, the source assigned to the current item in the list may be affected by the source of the previous item. A different explanation of the clustering is that subjects recall items from multiple sources, but their decision of which item to use is biased by their earlier decisions. The clustering of the recall items according to communication source is consistent with all of these interpretations. Recall lists provide no information to allow differentiation among the processes described above, but they do indicate a relationship between the sequence of items and communication source.

In this paper, we propose a test statistic for runs with multinomial data when the mean number of runs under randomness must be estimated from the data. Our statistic performs much better than a statistic derived using the distributional results of Mood (1940), which do not account for estimation.

## REFERENCES

Collins, A. M. and Loftus, E. F. (1975), "A spreading activation theory of semantic processing", *Psychological Review*, 82, 407-428.

Gauss-Version 1.49B (1986), Edlefsen, L. E. and Jones, S. D., Aptech Systems, Inc., Kent, WA.

Johnson, N. L. and Kotz, S. (1969), *Distributions in Statistics: Discrete Distributions*, New York: John Wiley and Sons.

Koppen, M. G. M. and Verhelst, N. D. (1986), "The exact runs test and some large sample approximations", *British Journal of Mathematical and Statistical Psychology*, 39, 168-182.

Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*, San Francisco: Holden-Day.

Lehmann, E. L. (1983), *Theory of Point Estimation*, New York: John Wiley and Sons.

Mood, A. M. (1940), "The Distribution Theory of Runs", *Annals of Mathematical Statistics*, 11, 367-392.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974), *Introduction to the Theory of Statistics* (3rd edition), New York: McGraw-Hill Book Co.

Murphy, M. and Puff, C. (1982), "Free Recall: Basic Methodology and Analysis," in *Handbook of Research Methods in Human Memory and Cognition*, ed. C. Puff, New York: Academic Press, pp. 99-128.

Pellegrino, J. W. and Hubert, L. J. (1982), "The Analysis of Organization and Structure in Free Recall," in *Handbook of Research Methods in Human Memory and Cognition*, ed. C. Puff, New York: Academic Press, pp. 129-172.

Robertson, C. (1985), "On the multiple runs distribution and its use in the measurement of categorical clustering", *British Journal of Mathematical and Statistical Psychology*, 38, 11-19.

Schwager, S. J. (1983), "Run probabilities in sequences of Markov-dependent trials", *Journal* *of the American Statistical Association*, 78, 168-175.

Shapiro, M. A. (1987), The Influence of Communication-Source Coded Memory Traces on World View, (Dissertation), University of Wisconsin-Madison.

Shapiro, M. A. (1988), "The Influence of Communication-Source Coded Memory Traces on World View", Proceedings of the Mass Communication Division of the International Communication Association, New Orleans.

Shaughnessy, P.W. (1981), "Multiple Runs Distributions: Recurrences and Critical Values", *Journal of the American Statistical Association*, 76, 732-736.

## APPENDIX 1

In this appendix we list the formulas for $\mathrm{Var}(r-\hat{\mu}_r)$ for ease of use in practice. $\mathrm{Var}(r-\hat{\mu}_r)$ is equal to $\mathrm{Var}(r) + \mathrm{Var}(\hat{\mu}_r) - 2\,\mathrm{Cov}(r, \hat{\mu}_r)$. Var(r) is given in (2.2).

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_r) &= \frac{1}{n^2}\Big(\sum_i \mathrm{Var}(n_i^2) + \sum\sum_{k\neq l} \mathrm{Cov}(n_k^2, n_l^2\,)\Big) \\
&= \frac{1}{n^2}\Bigg\{\sum_i\,[n(n-1)(n-2)(n-3)p_i^4 + 6n(n-1)(n-2)p_i^3 + 7n(n-1)p_i^2 + np_i] \\
&\quad - \sum_i\Big(n^2(n-1)^2p_i^4 + 2n^2(n-1)p_i^3 + n^2p_i^2\Big) \\
&\quad + \sum\sum_{k\neq l}[n(n-1)(n-2)(n-3)p_k^2p_l^2 + n(n-1)(n-2)(p_k^2p_l + p_kp_l^2) \\
&\quad + n(n-1)\;p_kp_l] \\
&\quad - \sum\sum_{k\neq l}\Big(n^2(n-1)^2\;p_k^2p_l^2 + n^2(n-1)\;(p_k^2p_l + p_kp_l^2) + n^2p_kp_l\Big)\Bigg\}\,.
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathrm{Cov}(r,\hat{\mu}_r) &= \Big[-\sum_i(n(n-1)p_i^2 + np_i)\;\Big]\Big[\frac{1}{n} + (\sum_i p_i^2)(1 - 1/n)\Big] \\
&\quad + \frac{1}{n^2}\Big(\sum_i\,[n(n-1)(n-2)(n-3)p_i^4 + 6n(n-1)(n-2)p_i^3 + 7n(n-1)p_i^2 + np_i] \\
&\quad + \sum\sum_{k\neq l}[n(n-1)(n-2)(n-3)p_k^2p_l^2 + n(n-1)(n-2)(p_k^2p_l + p_kp_l^2) \\
&\quad + n(n-1)\;p_kp_l]\Big).
\end{aligned}$$

## APPENDIX 2

All simulations were written in the matrix language GAUSS and were run on an IBM PC-AT computer. All random number generation was performed using the built-in random number generator in GAUSS, RNDU. Common random numbers were used to compare the means, standard deviations and tail probabilities of five statistics:

$$\frac{r-\mu_r}{\sigma_r}\ ,\ \frac{r-\mu_r}{\hat{\sigma}_r}\ ,\ \frac{r-\hat{\mu}_r}{\sigma_r}\ ,\ \frac{r-\hat{\mu}_r}{\hat{\sigma}_r}\ \text{and}\ \frac{r-\hat{\mu}_r}{SE(r-\hat{\mu}_r)}\ .$$

1500 replications were run for each n and p in order to achieve the desired accuracy in estimating the nominal $\alpha = .05$ tail probabilities. $SE(r-\hat{\mu}_r)$ and $\hat{\sigma}_r$ are zero if $\hat{p}_i$ is one for some i. In such cases, the test was declared to reject at all significance levels; however the standard deviation was not recorded. This occurred in the simulations only for n =10 for vector $p_3$. For that situation it occurred 3 times in 1500 replications.