Errors by macaques performing trial-unique delayed matching-to-sample: A function of memory per se or of distinguishing between the subsequent choices?

JAMES L. RINGO and ROBERT W. DOTY
Center for Brain Research, University of Rochester Medical Center, Rochester, New York

In trial-unique delayed matching-to-sample the animal must (1) remember the item given as the sample, and (2) subsequently distinguish it from a second item to make a match. With rather short delays, even very well trained monkeys continue to make errors on this task. The question is whether these errors arise as a consequence of poor memory per se, or whether inadequate memory makes the sample difficult to distinguish from the alternative. This question was examined using pairs of items presented on multiple occasions. The role of sample was systematically interchanged between the items forming a pair. Errors were found to be well correlated between trials in which the same item served as sample, but essentially uncorrelated when those trials were compared with trials having the other item as the sample. For example, if A and B were paired items, results of trials in which A was selected as the sample were well correlated with other trials using A as the sample (run on other days); however, trials using A as the sample were uncorrelated with trials using B as the sample, even though the comparison pair (A and B presented together) was identical. These results suggest that the monkeys' errors are not dependent on the distinguishability of the comparison pair, despite a "faded" memory on which to base the distinction.

In the delayed matching-to-sample (DMTS) task, an item is presented to a subject and then removed; following a delay, a pair of items, which includes the originally presented single item, is presented to the subject, whose task now is to choose which of the pair "matches" the originally presented item. In the trial-unique version of this task, items are used in one trial and then are not reused in that session. The trial-unique DMTS task and the related trial-unique delayed nonmatching-to-sample task are important tests in the behavioral analysis of animal memory. Their development (Harlow, 1944; Hayes & Thompson, 1953; Mishkin & Delacour, 1975) has not only provided a ready supply of quickly learned tasks, but has allowed for the extensive testing of event memory in monkeys (e.g., following temporal lobe lesions; Horel & Pytko, 1982; Mishkin, 1982). As in all such tests, there is the inevitable problem that other features of the task, such as sensory or motivational difficulties, may interfere with a clear measure of memory. This problem is particularly troublesome in work with animals, in which one cannot be sure of how the problem is being ap-

proached and must thus be cautious even in assuming the "obvious."

Because, on trial-unique DMTS, increasing the time lag between the removal of the sample and the presentation of the comparison pair (i.e., the delay period) leads to more incorrect choices, there is a strong presumption that memory is involved. As one can also reduce performance by selecting items that are arbitrarily similar, perceptual aspects of the task could also be a limiting factor. A commonsense approach to the analysis of DMTS performance might be to suppose that the sample is recorded and during the delay becomes less distinct. If the nonmatching item in the comparison pair is similar in appearance to the matching item, it might prove more difficult to differentiate it from the indistinct memory of the sample, whereas an item that could be more easily distinguished would cause less difficulty. Thus, differentiation based on "faded" memory could play a role. For example, clear photographs of a horse and a goat might be perfectly discriminable under 0-sec delay conditions, yet be similar enough to cause difficulty after a delay (in which the memory of whichever was the sample might fade). In other words, if an animal makes an error on a DMTS trial, then the reasonable presumption is that it forgot the sample. However, it may not have completely forgotten the sample, but may have forgotten just enough so that it can no longer distinguish between the two choice items. Thus, similarity of items that are nonetheless highly discriminable may play a role in DMTS performance. This paper

212

addresses the question of whether in a trial-unique DMTS task (using a wide variety of stimuli and a relatively short delay), errors are primarily due to inadequate memory, difficulty in distinguishing between partly similar items, or both. Because human subjects have no difficulty with either the discriminations or the delays in the ordinary trial-unique DMTS used on monkeys, one's intuition may be of little use in solving this problem.

To choose the correct match in the DMTS task, the animal must not only remember the sample but must also differentiate between it and its mate, despite possible fading of the sample memory. For example, if item A is the sample, A is shown to the animal and removed; after a delay, the animal is asked to pick A from the two choice items, A and B. If items A and B are at all similar, then even if some of A is well remembered, it may be that key features of A are not remembered (or retrieved) or that the entire memory of A is indistinct enough to make a distinction problematical. If the similarity between the items of the comparison pair is the major limiting factor, then the particular two items chosen for pairing and presentation in each DMTS trial would be critical; it would not matter which of the two was used as the sample. To assess the importance of the similarity of the two items in each pair, we have examined the correlation of errors between trials, using the same pair of items in each trial, but using A as the sample in one trial and B in the other. These trials were run on alternate days in a trial-unique DMTS task in which no items were repeated on one day. If item similarity is a major source of difficulty, then errors should occur regardless of which item of the pair serves as the sample. Thus, the correlation of errors for trials using the same comparison pair ought to be high, regardless of which item happens to be the sample. On the other hand, if the difficulty is primarily a matter of remembering the sample, then there should be little correlation between the error rate when item A serves as the sample and when item B so serves.

METHOD

Subjects

Three adult male *Macaca nemestrina* were used in this study. These animals had been used in a previous study and had been through various surgical procedures: DND had a sectioned optic chiasm, HUD had a sectioned optic chiasm and anterior commissure, and RHD had a partly sectioned optic chiasm and fully sectioned corpus callosum.

Apparatus

A more complete description of many features of the apparatus can be found in Overman and Doty (1980). Briefly, the monkeys were trained and tested in a sound-reducing chamber equipped with a one-way observation window. The monkeys were seated in a restraining chair that allowed free arm movement and were positioned 25 cm in front of three vertically aligned plastic rear-projection panels (each subtending 20° horizontally and 14° vertically). Three Kodak Carousel projectors, equipped with 180-mm lenses and extension tubes, projected standard 35-mm slides as high-resolution 5.5 × 7 cm images. First the central panel was illuminated for a few seconds with a particular image (the sample), and the monkey

was rewarded with juice (typically 0.5 ml) for pressing it. The image was then extinguished and a 2- to 10-sec delay ensued. Images then appeared on the upper and lower panels, one of which (the match) was identical to that just presented, and the monkey was rewarded with 2-5 ml of juice for pressing the panel on which that image appeared. Choice of the other image (the alternative) could produce a puff of air directed at the top of the monkey's head, a loud toot of a horn, and/or a 20- to 50-sec time-out before presentation of the next sample (normally a 5-sec intertrial interval).

Viewing of the image was restricted to one eye (and initial input to one hemisphere in these split-chiasm animals) by rotary shutters mounted on a lightweight mask worn by the monkeys. Since the correlations measured in this study were found to be the same regardless of which eye viewed the material, data from the various eye combinations were combined.

Materials

Five groups of slides were used: Group 1, complex colored objects—these were slides of various objects found around the laboratory or home; Group 2, black-and-white objects—these were black-and-white duplicates of the slides in Group 1; Group 3, complex colored objects not in Group 1 (and hence not duplicated in Group 2); Group 4, hue slides—slides of homogeneous color which, when displayed to the monkey, appear as featureless rectangles of color; Group 5, counting slides—pairs of slides, one showing a few and one showing many, of identical objects, for example, coins or fruit. Slides shown in any one session were from only one group. In order to somewhat equalize performance, the Group 5 material was tested with only a 2-sec delay. The Group 4 material was tested with a 5-sec delay for HUD and 10-sec delays for the others. All other testing was with a 10-sec delay.

Procedure

A session consisted of 45 or 50 trials. Images were paired for each trial and one image served as the sample. In the next session, the same images were paired but now the other image served as the sample, and so on. So, for example, if the paired items for the first trial were A and B, then on Trial 1 in one session, A was the sample (and correct choice) and on Trial 1 in the next session, B served as the sample. For convenience, we will call these two sequences X and Y. Sequence Y counterbalances sequence X and prevents discrimination learning. The slides were displayed so that the correct image of the comparison pair (the match) appeared in a pseudorandom order on the top and bottom panels, limited to three consecutive correct on either panel and restricted to produce equal numbers of correct top and bottom panel displays (22-23 for 45 total trials). From 6 to 16 sessions of one sequence were run (always with an equal number of sessions using the counterbalancing sequence). The slides chosen to make up a trial were selected with only ordinary care to exclude pairs that appeared to be particularly difficult to discriminate. Stimuli were not reused during a session and so in this sense were trial-unique, but obviously, stimuli were repeated when a sequence or its counterbalancing sequence was repeated. Examination for time trends in the animals' performance as they gained experience with a particular sequence (and its counterbalanced version) showed no evidence that the animals learned the sequences or fragments thereof.

Analysis

The monkeys' responses were counted as 1 if in error and 0 if correct. The few (< 1%) failures to respond were treated as wrong responses. Overall, the animals performed fairly well. No animal performed below 75% correct with any material. The average of the 3 animals' scores on each of the five groups of material was 92%, 89%, 90%, 83%, and 84% (Groups 1 to 5, respectively). Each session produced an ordered series of error counts 45 or 50 items long. Averaging between identical sessions was introduced

in order to reduce noise. Averaging was done by adding all the errors made on Trial 1 in sessions to be included in the average and dividing by the number of sessions, and so on for each trial. Pearson product-moment correlations were determined between averages. Four averages were formed from the sessions run using material from one group. Two averages (X1 and X2) were formed from the sessions with sequence X and two averages (Y1 and Y2) were formed from the sessions run with the counterbalancing sequence Y. If on Day 1 a monkey was tested on sequence X, Day 2 testing would be on the counterbalancing sequence Y, and so on for 16 sessions. Then four averages would be formed, two from the first session type, X1 (Days 1, 5, 9, and 13) and X2 (Days 3, 7, 11, and 15), and two from the counterbalancing sequence, Y1 (Days 2, 6, 10, and 14) and Y2 (Days 4, 8, 12, and 16). Two averages were formed from each sequence type (e.g., X1 and X2 from the X sessions) in order to allow for correlations between averages of sessions of the same sequence. Correlations between counterbalancing sequences were performed on averages of only half the sessions in order to keep the same amount of noise reduction (from averaging) as in the correlations made within types. All available sessions were brought into the averaging; there were generally four or more sessions per average, but for RHD and DND on Group 1 material there were fewer sessions available. Two correlations were performed between results from identical sessions: X1 to X2 and Y1 to Y2. The two resultant correlation coefficients were averaged to produce one correlation value between identical sessions for each animal working on each group of material. This value is termed $r(X,X; Y,Y)$. Similarly, four correlations were performed between counterbalanced sessions: X1 to Y1, X1 to Y2, X2 to Y1, and X2 to Y2. The four resultant correlation coefficients were averaged to produce one correlation value between counterbalanced sessions for each animal working on each type of material. This value is termed $r(X,Y)$.

For each correlation calculated from the data averages, a computer simulation was run to estimate the distribution of correlation coefficients due to chance (i.e., in the uncorrelated case). This was done by randomly shuffling one of the two series (ordered by trial number) to produce a random reordering, thus forming an alignment uncorrelated with the other series. Then a correlation coefficient was calculated, another random reshuffle was done and another coefficient found, and so forth. An outline of the distribution for the uncorrelated case was produced by finding 1,000 correlation coefficients from random reshuffles. This procedure is not primarily intended for significance testing (although significance can be assessed from the extremes of the distribution). The main point is to show that these correlation coefficients are reasonably "well behaved" so that we may apply our ordinary understanding of correlation to the problem at hand.

RESULTS

The basic result is that averages from sessions of the same sequence type (X,X or Y,Y) were well correlated, whereas correlations between averages from counterbalancing sequences (X,Y) were very small. Averaging of identical sessions allows a much clearer separation between the two types of correlations. The effect of averaging can be seen in Table 1, where the number of sessions averaged is 4, 3, 2, or 1. The averaging clearly reveals a correlation between identical sessions that is obscured by session-to-session noise. In other words, the average of the correlations between single sessions is low for both X,X (or Y,Y) and X,Y types. However, the correlations between averages (i.e., when four sessions

Table 1
Correlation Coefficients Between Averages From Identical or Counterbalanced Session Types with 1,2,3, or 4 Sessions Averaged, Group 3 Material

Animal	Number of Sessions	Average $r(X,Y)$	Average $r(X,X; Y,Y)$
HUD	1	0.03	0.12
	2	0.01	0.20
	3	0.06	0.23
	4	0.04	0.29
RHD	1	−0.06	0.23
	2	−0.10	0.41
	3	−0.14	0.44
	4	−0.13	0.58
DND	1	−0.03	0.18
	2	−0.05	0.29
	3	0.00	0.50
	4	−0.09	0.57

Note—All sessions are included except for the case of the three-session average, in which the first sessions of both A and B types were dropped from the four-session average.

are averaged before correlations) reveal an underlying correlation for the X,X (or Y,Y) types but none for X,Y types.

Figure 1 shows the distribution of the randomly generated correlation coefficients (the density functions). This figure shows the six correlations from HUD running on Group 1 material: $r(X1,X2)$, $r(X1,Y1)$, $r(X1,Y2)$, $r(X2,Y1)$, $r(X2,Y2)$, and $r(Y1,Y2)$. The arrow in each density function marks the correlation coefficient of the actual data. The points in this figure are the number of randomly generated correlation coefficients found out of 1,000 in each bin with 0.1 width. The bins were positioned so that the median is on the border of the two central bins.

With the exception of DND working on Group 1 material, which generated too few errors in too few sessions to produce a smooth distribution (but was included for completeness), the random distributions found from reshuffling the data from all animals on all groups showed similar distributions (similar in median, range, and shape). For example, in all the random distributions obtained, the 10th smallest value in each distribution (out of 1,000, hence the 1% point) ranged from −0.19 to −0.35, whereas the range of values found at the 99% level (the 10th largest value of the 1,000) was 0.31 to 0.46. The median ranged from −0.047 to 0.0059. The constructing of the random distributions demonstrates that the underlying distributions of the correlation coefficients between sequences of the same type are apparently identical to the random distribution of correlation coefficients between session averages and counterbalancing session averages, so that the correlations obtained from the data are directly comparable.

The correlation coefficients calculated between averages formed from identical sequences are listed in the right-hand column of Table 2 as $r(X,X; Y,Y)$. This is the average of two correlations, one from X1 and X2 and the
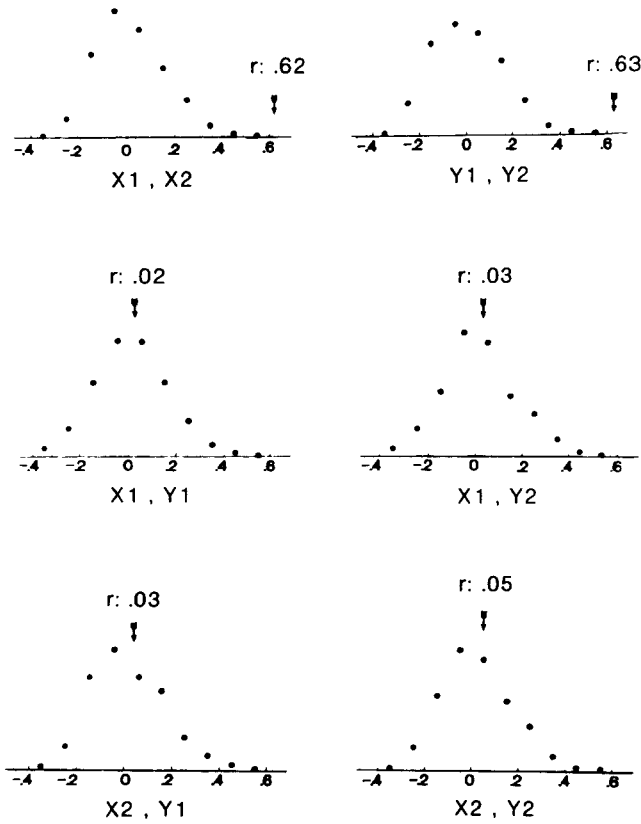
Figure 1. Probability density functions of the correlation coefficients generated from random and uncorrelated alignments of averaged data from animal HUD on Group 1 material. The actual correlation coefficient in each case is shown by the arrow. The correlation coefficients found between the averages X1 and X2 and between Y1 and Y2 are larger than any of those produced by their respective 1,000 random shuffles. The correlation coefficients between averages from counterbalancing sessions (X1,Y1; X1,Y2; X2,Y1; X2,Y2) are all near the middle of the distributions and are consistent with uncorrelated behavior.

other from Y1 and Y2. The grand average of all the correlations between identical sequences is 0.48. In strong contrast to this result is the lack of correlation between averages formed from a sequence and its counterbalanced sequence (X,Y). Here, despite the exact duplication, trial by trial, of the pair of images making up the comparison pair, the correlations are very small. The grand average of all the correlations between averages from X and Y sequences is −0.05. These correlation coefficients are listed in the column of Table 2 labeled $r(X,Y)$. Statistical testing confirms what is apparent from inspection of Table 2 that the correlations between identical sessions are highly significantly greater than the correlations between counterbalancing sessions ($p < .001$, sign test).

Subtotals collected for each animal show consistent results. None of the animals' totals is far from the grand average. If the data are collapsed across materials and summed separately for each animal, the correlations across different session types [$r(X,Y)$] are still low (RHD,

−.09; HUD, −.04; DND, +.01). The correlations between identical session types [$r(X,X;Y,Y)$] are much stronger (RHD, .51; HUD, .50; DND, .42).

A similar examination shows that the results are consistent over the various types of visual material employed in this study. That is, if the data are collapsed across the animals, but separately for each type of material, the correlations between different session types [$r(X,Y)$] are still low (Group 1, colored objects, −.07; Group 2, black-and-white objects, +.09; Group 3, colored objects, −.06; Group 4, hue, −.06; Group 5, few/many, −.13). The correlations between identical session types [$r(X,X; Y,Y)$] are much stronger (Group 1, .48; Group 2, .55; Group 3, .48; Group 4, .45; Group 5, .45).

DISCUSSION

As can be seen in Table 2, most of the correlations between a sequence and its counterbalancing sequence were small even after averaging. Correlations between averages formed from the same sequence types, however, are significantly larger. This indicates that on identical sequences errors tend to be made on the same test images (trials), whereas between sequences and their counterbalancing sequences, errors do not tend to be on the same test pair. That is, a monkey running on two X-type sessions, or two Y-type sessions, will tend to make errors on the same trials in each session (i.e., the same test item). But a monkey running on an X-type session and a Y-type session will tend not to produce errors on the same trials, even though the pair of images is the same and only the member serving as sample is different. The low correlations between sequences and their counterbalancing sequences imply that errors are caused by something not

Table 2

Correlations Between Averages of Sessions Using Sequences and Their Counterbalancing Sequences $r(X,Y)$, and Correlations Between Averages Formed From Sessions Using the Same Sequences $r(X,X; Y,Y)$

Material	Animal	Number of Sessions	$r(X,Y)$	$r(X,X; Y,Y)$
Group 1:	RHD	12	−.19	.40†
Colored objects	HUD	16	.02	.62‡
	DND	8	−.03	.43†
Group 2:	RHD	24	−.07	.66‡
B & W objects	HUD	24	.01	.53‡
	DND	32	.32*	.45†
Group 3:	RHD	16	−.13	.58‡
Colored objects	HUD	16	.04	.29*
	DND	16	−.09	.57‡
Group 4:	RHD	16	.04	.44†
Hue	HUD	16	−.15	.45†
Group 5:	RHD	16	−.09	.49†
Few/many	HUD	16	−.14	.61‡
	DND	16	−.17	.24
Grand Averages			−.05	.48

Note—Data from DND are absent from Group 4, since he was not run on that group. *$p < .05$. †$p < .01$. ‡$p < .001$. (Probabilities estimated from the simulation.)

shared by the two sequence types. The obvious candidate here is the sample, since the delay and the comparison pair are unchanged between a sequence and its counterbalancing sequence. Apparently some feature or characteristic of the sample makes a trial easy or difficult. It is of interest that, as might to some degree be expected, the errors made by each animal in the present experiment were much better correlated to those of that individual than to those of the others, although the overall correlation across individuals was significant, but low. In other words, although there was a tendency for all monkeys to find certain items to be more difficult than others, each monkey had its own unique tendency to err on certain items.

The major result, which is apparent in the grand averages of Table 2, is also evident in the data for each animal (when the data are collapsed across material types) and for each material type (when the data are collapsed across animals). Since each of the three animals produced results reasonably near the grand averages, these results are apparently consistent in the face of individual variation. Similarly, the consistency of results across types of material indicates that the main result is not dependent on the peculiarities of a single type of material. The agreement among the results for the first three groups of material is unsurprising, because these are similar materials. Groups 1 and 3 were essentially just different collections of items of a comparable nature. Group 2 might well have been expected to produce similar results, because these were the black-and-white versions of the photographs used in Group 1. The last two groups, however, were greatly different. Group 4 was composed of pure hue slides, without contour or form cues, and Group 5 was composed of pairs in which the slides differed in the number of examples of an object, although the object shown was the same for both slides in one pair. A somewhat different strategy or analysis may have been required of the animals confronted with these last two groups, compared to the others, yet the correlations remain robust.

The lack of correlation between trials in which the comparison pair was maintained while the member of the pair serving as sample was switched implies that the errors made by the monkeys were not caused by difficulty in differentiating the pair of images from which the match was to be chosen (since this would have produced positive correlations). This result is encouraging for those using the visual DMTS as a memory task, because the data are consistent with a system limited by memory for the sample (this despite any visual difficulties introduced by the optic chiasm section in our monkeys). Wright and Sands (1981) have shown that the pigeon will often make its match in a DMTS task without even observing the alternative. This result, too, is consistent with a memory limit, rather than reflecting any problem with the differentiation of the comparison pair. The possibility that stimulus preference could be involved is raised by the positive correlation found between repetitions of the same trials.

However, stimulus preference would lead to negative correlations between counterbalanced session types, because use of the preferred item as the sample would lead to a disproportionate number of correct responses whereas use of the nonpreferred item as the sample would lead to a disproportionate number of wrong responses. Stimulus preference strong enough to lead to a strong positive correlation between identical trials should lead to a strong negative correlation between counterbalanced sessions. Because that was not found to be the case, it seems that stimulus preference exerted no important influence here.

For a DMTS task in which only a few pairs, or only a single pair, of objects are used for all trials, there is much support for the idea that the subjects are engaging in a temporal discrimination (D'Amato, 1973; Mason & Wilson, 1974; Worsham, 1975). That is, the subjects have seen the stimuli many times and need to discriminate which stimulus of the comparison pair they have seen most recently. In the trial-unique DMTS task, on the other hand, items are not repeated within a session but are often reused on following sessions. A temporal discrimination between the two items of the comparison pair in this case might still be a limiting factor on the subjects' abilities. It is difficult, however, to suppose that our monkeys were engaging in a temporal discrimination, because such a discrimination presumably would be harder or easier depending on the comparison pair used. That, in turn, would have produced a correlation between trials having the same comparison pair even with alternate samples, a correlation that was not found. It may well be that the temporal discrimination has been made trivial by the trial-unique procedure and no longer limits the monkeys' performance. This, in fact, is just the logic behind the use of trial-unique DMTS (Mishkin & Delacour, 1975).

Our results are, of course, applicable only to the range of material used in our study, and only on the trial-unique DMTS task. Presumably they would also apply in the case of nonmatching-to-sample. However, with a non-trial-unique DMTS task, Devine, Burke, and Rohack (1979) found that when the two choices for the match were colors and/or shapes, their monkeys performed better if the matches were from different categories of materials; that is, in this extreme case, the difficulty of the discrimination did matter. Perhaps in that situation the difference in the comparison pair mattered not so much because it eased a difficult discrimination as because it altered the nature of the memory problem, so that all the monkey needed to remember was whether the sample was a color or a form, not which color or form. Similarly, on a visual discrimination task (e.g., Polidora, 1966), or a task in which the match pair are nearly identical (e.g., in psychophysical measurements), the match discrimination will be important. Nonetheless, when the items used on a DMTS task are from the same category and, in the judgment of the experimenters, reasonably discriminable, the problem of differentiating between the choice pair, even on the basis of a faded memory, can be reduced to insignificance.

REFERENCES

D'AMATO, M. R. (1973). Delayed matching and short-term memory in monkeys. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 7, pp. 227-269). New York: Academic Press.

DEVINE, J.V., BURKE, M. W., & ROHACK, J. J. (1979). Stimulus similarity and order as factors in visual short-term memory in nonhuman primates. *Journal of Experimental Psychology: Animal Behavior Processes, 5,* 335-354.

HARLOW, H. F. (1944). Studies in discrimination learning by monkeys: II. Discrimination learning without primary reinforcement. *Journal of General Psychology, 30,* 13-21.

HAYES, K., & THOMPSON, R. (1953). Nonspatial delayed response to trial-unique stimuli in sophisticated chimpanzees. *Journal of Comparative & Physiological Psychology, 46,* 498-500.

HOREL, J. A., & PYTKO, D. E. (1982). Behavioral effects of local cooling in temporal lobe of monkeys. *Journal of Neurophysiology, 47,* 11-22.

MASON, M., & WILSON, M. (1974). Temporal differentiation and recognition memory for visual stimuli in rhesus monkeys. *Journal of Experimental Psychology, 103,* 383-390.

MISHKIN, M. (1982). A memory system in the monkey. *Philosophical Transactions of the Royal Society of London,* **B298,** 85-95.

MISHKIN, M., & DELACOUR, J. (1975). An analysis of short-term visual memory in the monkey. *Journal of Experimental Psychology: Animal Behavior Processes,* 1, 326-334.

OVERMAN, W. H., & DOTY, R. W. (1980). Prolonged visual memory in macaques and man. *Neuroscience, 5,* 1825-1831.

POLIDORA, V. J. (1966). Stimulus correlates of visual pattern discrimination by monkeys: Multidimensional analyses. *Perception & Psychophysics,* 1, 405-414.

WORSHAM, R. W. (1975). Temporal discrimination factors in the delayed matching-to-sample task in monkeys. *Animal Learning & Behavior,* 3, 93-97.

WRIGHT, A. A., & SANDS, S. F. (1981). A model of detection and decision processes during matching to sample by pigeons: Performance with 88 different wavelengths in delayed and simultaneous matching tasks. *Journal of Experimental Psychology: Animal Behavior Processes,* 7, 191-216.