Animal Learning & Behavior
1992, 20 (3), 240-252

Primacy, recency, and the variability of data in studies of animals' working memory

E. A. GAFFAN
University of Reading, Reading, England

The performance measures in many experiments on animal memory are expected to have an underlying binomial distribution, with additional variance contributed, for example, by between-subject differences. This paper examines whether the data from published studies of serial position effects (primacy and recency) in animals' working memory conform to that expectation. In most cases, the variance, when it can be estimated, is consistent with those statistical assumptions, but in certain studies, it is significantly smaller than expected. This is usually a sign of faulty procedure or analysis, and possible causes are discussed. The conclusion is that much of the evidence for primacy in animals is unsatisfactory, on statistical or other methodological grounds. The analytic approach outlined here might usefully be applied to detect potential problems with other experiments of a similar type, especially when manually operated apparatus is employed, and to improve their statistical power.

In this paper, I describe a simple statistical approach and apply it to studies of serial position effects in animals' working memory. I therefore review and reevaluate previous investigations of serial position effects, as well as discuss how the statistical principles could be more widely exploited.

In many experiments, animals have been required to remember sequences (lists) of items, either spatial (e.g., arms of a radial maze) or nonspatial (e.g., objects, pictures, patterns). I concentrate here on working memory paradigms, in which different lists are presented on different trials and the animals' memory for the most recent list is assessed. It has been claimed that animals' performance under these conditions often resembles that of people in analogous tests of recall or recognition of items from lists of words or pictures (Wright & Watkins, 1987). In particular, animals are said to show, under appropriate conditions, both recency (superior memory for items nearer the end of the list) and primacy (better memory for items early in the list, rather than in the middle). (See the references below.)

As Hitch (1983, 1985) has commented, primacy and recency have rather different status. Recency effects in working memory are readily obtained from both animal and human subjects, with whatever type of material is to be remembered, and Hitch suggests that they reflect similar processes (simple temporal decay, retroactive interference) in all of the species that have been tested. Consistent with the principle of temporal decay is the finding that items from the end of a list lose their advantage if sufficient time elapses before recall, as will be illustrated below.

However, primacy effects are more elusive. Hitch argues that primacy in people most probably reflects rehearsal and elaborated coding, particularly though not exclusively available for verbal material (although other mechanisms have been proposed—cf. Wright, 1989). If so, genuine primacy effects are a priori less likely to be found in animals. Indeed, D. Gaffan (1983) contended that apparent primacy effects in animal memory (Roberts & Kraemer, 1981; Sands & Wright, 1980) were artifactual; they were evident only in procedures in which the subject made an observing response before the first item of the list was presented. Primacy might simply reflect higher attention (i.e., more efficient perceptual input) for the first item of the list, rather than differential memory processing.

Many new data have appeared since that time, and in this paper, I will reconsider the status of serial position effects in the light of that. The first section below is a summary of recent claims about determinants of serial position effects in animals. After that, my main purpose will be to present a partial reanalysis of some of the data in order to evaluate their statistical properties—in particular, whether their variability conforms to certain reasonable assumptions. The third section is a discussion of the implications of that reanalysis, and in the final section, I make some general recommendations for research of this type as well as reassess the evidence for primacy and recency effects.

DETERMINANTS OF PRIMACY AND RECENCY IN SPATIAL AND NONSPATIAL MEMORY

In a series of papers, Kesner and colleagues (see references) have described both primacy and recency effects

I thank D. Gaffan, P. T. Smith, V. M. LoLordo, and anonymous reviewers for helpful contributions to this paper. Correspondence should be sent to E. A. Gaffan, Department of Psychology, University of Reading, Reading RG6 2AL, England (e-mail: sysgaffn@uk.ac.rdg.susssys1).

in rats' memory for lists of spatial locations—namely, arms of a radial maze. They have tested both *item* memory—when subjects must discriminate between an item that has occurred in the recently presented list and one that has not—and *order* memory, when subjects are presented with two items, both of which have occurred in the recent list, and must distinguish which one occurred earlier.

Kesner has concluded that both primacy and recency are consistently observed in rats' order memory. For example, if presented with a list of eight maze arms that they must enter in a sequence determined by the experimenter, rats discriminate well the order of the arms that occur first and second in the list, and that of the arms that occur seventh and eighth, but they are poor at discriminating the order of a pair of arms that occur in the middle of the list, such as the fourth and fifth. U-shaped serial position curves, showing good performance on early and late parts of the list but near-chance performance on the middle, have been obtained with the use of both four- and eight-item lists (Kesner, Adelstein, & Crutcher, 1987; Kesner, Crutcher, & Measom, 1986; Kesner & Gray, 1989; Kesner & Holbrook, 1987; Kesner, Measom, Forsman, & Holbrook, 1984; Kesner & Novak, 1982).

In the case of item memory, Kesner suggests that recency is consistently obtained, but whether or not primacy occurs depends on the type of discrimination that the rats learn. When a rat is tested with a choice between two arms—one familiar (i.e., entered during presentation of the list) and one relatively novel—it may be rewarded for choosing the novel one (the shift or win-shift rule) or for choosing the familiar one (the stay or win-stay rule). Both in Kesner's studies and in previous similar experiments done by others, rats trained under a shift contingency have shown no primacy, only recency (Cook, Brown, & Riley, 1985; DiMattia & Kesner, 1984; W. A. Roberts & Smythe, 1979). Their preference for a novel over a familiar arm is more accurate the later in the list—that is, the more recently that the familiar arm has occurred. However, Kesner reports that rats trained to follow a stay rule show both primacy and recency. Their preference for a familiar over a novel arm is strong when the familiar arm occurs either early or late in the list, but weaker when the familiar arm is from the middle of the list. Such U-shaped serial position curves following stay training have been presented by DiMattia and Kesner (1984); Kesner, Adelstein, and Crutcher (1989); Kesner, Crutcher, and Beers (1988); Kesner and Gray (1989); and Kesner and Holbrook (1987).

Other laboratories do not appear to have reported any replications of findings of U-shaped (primacy plus recency) serial position curves for rats' spatial memory under the conditions that Kesner and colleagues state to be necessary—namely, order memory training, or stay contingency item memory testing. Dale (1987), who used an analogous paradigm with human subjects, observed both primacy and recency in order memory, but no consistent primacy in item memory. Maki, Beatty, and Clouse (1984) mentioned that, in unpublished experiments, they could not obtain similar serial position effects for rats' order memory, and in one study of rats' item memory done by DiMattia and Kesner (1988), there was no evidence of primacy in the stay contingency group.

Lack of independent replication alone would not be enough to cast doubt on a large number of published studies, but several other recent findings imply that Kesner's conclusions should be modified. Bolhuis and van Kampen (1988, Experiment 2) showed that both primacy and recency could be observed in rats' spatial memory, despite using a shift and not a stay contingency. The retention interval was the critical variable; they found only recency when the test trial occurred immediately after the end of the list (as it did in all the studies reviewed above) but only primacy when the test was delayed for 16 min. Caunt (1990), in an unpublished experiment done with the same task and with an intermediate retention interval, observed both primacy and recency. A stay contingency, then, does not appear to be necessary.

The transition from recency to primacy as retention interval is lengthened is consistent with the results of other studies of various species' memory for nonspatial stimuli such as pictures, objects, or sounds. When memory was tested immediately after the end of the list, only Sands and Wright (1980) and W. A. Roberts and Kraemer (1981) observed primacy. Others found only recency, whether a single list item was probed (D. Gaffan, 1977; Macphail, 1980; W. A. Roberts & Kraemer, 1984; Shimp, 1976; Thompson & Herman, 1977; Wright, Santiago, Sands, Kendrick, & Cook, 1985) or the whole list was tested in reverse order (D. Gaffan & Weiskrantz, 1980).

If the whole list is immediately tested in its original order, not only is primacy absent, but the last items in the list lose their normal superiority in recall (Moss, Rosene, & Peters, 1988; Overman, McLain, Ormsby, & Brooks, 1983). The same-order testing procedure necessarily imposes a delay between presentation and test for every item, so the disappearance of "recency" under that condition implies that the effect depends on an item's being tested within a short time from its presentation, not on its having occurred near the end of a list. That conclusion is supported by the results of studies in which, as in Bolhuis and van Kampen (1988), single items have been tested at varying delays after the end of each list. As the retention interval increases, not only does the recency effect diminish as expected, but a clear primacy effect often appears (Wright et al., 1985). A variety of contingencies have been used in these nonspatial memory experiments—sometimes matching (analogous to spatial stay), sometimes nonmatching (like spatial shift), and sometimes neither, with the presentation instead of a single probe picture that either has or has not been included in the list, the subject being required to make one of two responses signifying "old" and "new," respectively. (Buchanan, Gill, & Braggio, 1981, observed both primacy and recency in a single chimpanzee, but that study is not rel-

evant here. The animal was sophisticated in an artificial language, "Yerkish," and the paradigm was free recall—i.e., reproduction of lists of Yerkish symbols. Neither the subject nor the procedure can validly be compared to those considered in this paper.)

Returning to D. Gaffan's (1983) suggestion that primacy, when present, reflects the attention-enhancing effect of an observing response at the start of each list, it should be noted that usually when primacy has been found in animals that have not been language-trained (e.g., W. A. Roberts & Kraemer, 1981; Wright et al., 1985), observing responses have been employed; and in Dale's (1987) spatial list paradigm for human subjects, the list was immediately preceded by a distinctive warning cue. By contrast, D. Gaffan and Weiskrantz (1980), who observed no primacy even when they imposed a delay between the end of the list and the test, did not require an observing response. However, the fact that primacy can, at least under some circumstances, be observed in rats' *spatial* memory implies that an observing response per se is not necessary, because radial maze trials are initiated by the experimenter and not by the rat.

In a recent paper, Reed, Chih-Ta, Aggleton, and Rawlins (1991) apparently extended that conclusion by reporting clear U-shaped serial position curves in rats' memory for nonspatial lists (distinctive goalboxes attached to arms of a Y-maze), again in the absence of any explicit observing response. However, E. A. Gaffan and D. Gaffan (in press) argue that Reed et al.'s data do not provide valid evidence for their conclusions. The variability is significantly too low, according to arguments to be outlined below. Indeed, the authors have subsequently acknowledged that the report should not be relied upon (Rawlins, Deacon, Chih-Ta, & Aggleton, in press; Reed, in press). This led me to ask whether some of the apparent inconsistencies in the literature on serial position effects have arisen in part from similarly unsatisfactory data. The next section examines the statistical properties of data from other published experiments.

VARIABILITY OF DATA IN STUDIES OF SERIAL POSITION EFFECTS

A common feature of many experiments on animal memory, including those mentioned hitherto, is that performance has been measured in terms of the accuracy of choice between two alternatives—novel versus familiar or new versus old in item memory, and earlier versus later in order memory. Unlike other common measures of behavior, such as response latency or magnitude, the probability of correct choice gives rise to a binomial distribution of scores, which has the notable property that its variance is predictable from its mean.

Specifically, by the binomial theorem (e.g., Hays, 1988), when there are T trials of a choice between two alternatives, with a probability P of making the correct choice, the variance of the number of correct choices should be $TP(1-P)$. It is assumed, of course, that successive trials are independent events. It is reasonable to assume independence in most studies of serial position effects; typically, the different serial positions are probed, one per trial, in quasirandom order, so there is no basis for sequential dependency within the set of trials over which one counts correct responses by any one subject at any one serial position.

Suppose N subjects are tested, each for T trials, on a particular serial position, and the number of correct choices made by each animal is counted. Suppose, for the moment, that the underlying value of P (probability of correct response) is the same for all animals. Then the variance among the animals' scores is expected to be $TP(1-P)$, and the standard error of the mean score is expected to be $\sqrt{[TP(1-P)]}/\sqrt{N}$. From now on, this set of simple assumptions and predictions will be referred to, for brevity's sake, as the *binomial model.*

There are of course other sources of variance beyond the intrinsic binomial variance. For example, if there are differences between animals in their skill at the task—that is, if they have different individual values of P—then the overall distribution of their scores will superimpose binomials with different P values, and be broader than a single binomial distribution whose P is the mean of the animals' individual Ps. In other words, between-subject variance in P will be added to fundamental binomial variance. Such between-subject variation might be especially likely when a task requires much skill or practice, or the animals' performance is affected by brain damage whose extent is variable. There are yet other sources of variance such as random fluctuations in subjects' attention, experimental procedure, or accuracy of data recording. So observed variance between animals might commonly be greater than the minimum expected from the binomial model alone.

However, it is also possible for observed variances to be smaller than expected. If the discrepancy is significant, as it was in the case of Reed et al.'s (1991) data (see below), this implies that some constraint or bias limits the variance of the scores below its natural minimum. Such a constraint could arise for a number of reasons, many of which are methodologically undesirable, such as inadequate randomization of conditions across animals, experimenter effects, or biases in selecting or recording data. Some of these could be present even if experiments were automated, but most are self-evidently more possible when types of manually operated apparatus, such as mazes or Wisconsin test boxes, are used. I will argue that excessively low variance implies the need for caution in interpreting experimental findings. It can also obviously inflate the risk of Type I error in statistical analysis.

I will now examine how well the observed variances of scores, in published studies of serial position effects and related phenomena, conform to what is expected from the binomial model. As indicated above, larger than expected variance could occur for many reasons, but smaller than expected variance is usually problematic.

To test discrepancies from expectation, one measures the ratios of observed to expected variance for separate data points (e.g., group performance at different serial

positions). Expected variance $TP(1-P)$ is computed from T, the number of trials per subject per data point, and P as estimated from the group mean probability of correct responses at that point. According to the binomial model, the ratios of observed to expected variance should cluster round 1. The statistical significance of deviations from 1 can be tested by the chi-square statistic.

If N samples are taken from a normally distributed population whose variance is σ^2, and the variance among the samples is s^2, then $(N-1)s^2/\sigma^2$ is distributed as chi-square with $N-1$ degrees of freedom (Hays, 1988). According to our model, the distribution of the number of correct responses by a group of animals is binomial and can be approximated by a normal distribution with the variance $TP(1-P)$ as previously defined. So the ratio of the observed variance to the expected variance in a set of such scores (s^2/σ^2) can be multiplied by $N-1$, where N is the number of subjects, and the result tested against chi-square tables.

As well as using the chi-square test in the familiar way to detect chi-square values that are improbably large, by using the $p < .05$ criterion, one can detect values that are too small by using the $p > .95$ criterion (cf. S. Roberts, 1987). Values as small as that criterion should occur only 5% of the time under the null hypothesis. These criteria correspond to the two types of case discussed above, in which variances are either larger or smaller than expected from the binomial model.

To permit this analysis, either individual subjects' scores must have been reported, or variance statistics such as standard errors, preferably for each data point separately. It is not usually possible in the present case to calculate within-cell variance from reported cell means and F or t statistics, because totally within-subject designs are used, so tests are based on subject × condition error terms, allowing no estimate of between-subject variance. The exception is the study by Dale (1987), who used a partly between-subject design and reported exact F ratios, so that the average within-cell variance could be estimated. Inevitably some of my variance estimates are inaccurate, either because of the need to measure graph lines, or because the graphs themselves may be erroneous. The overall conclusions, however, remain clear. Only strong and consistent deviations from expectation are of interest, and to detect these, extreme accuracy is not usually necessary. Whenever variance was at the margin of being significantly too low, measurements were repeated with the graphs more highly enlarged.

Not all reports of serial position effects provide the necessary information, so, to give grounds for comparison, I have also examined certain studies in which variance was estimated under other conditions (e.g., different retention intervals or stages of training) but with similar apparatuses and paradigms. The general aim is both to evaluate the degree to which the binomial model fits the data, and to assess whether variability differs across apparatuses, tasks, and subject populations in the way that I have conjectured. I will consider whether any sources of evidence for serial position effects other than that of Reed et al. (1991) show signs of excessively low variance.

The studies are grouped according to the type of memory test (item or order) and material (nonspatial or spatial), as well as according to whether the apparatus was automated or not.

Nonspatial Item Memory, Nonautomated Apparatus

The only published study of serial position effects that falls into this category is that of Reed et al. (1991). It is obviously important to evaluate their results against comparable ones. They used a nonmatching paradigm with rats, so I have examined several studies of nonmatching or matching performance, in which single items rather than lists were to be remembered. In three, rats were the subjects; in one, the same apparatus as that of Reed et al. was used, with complex goalboxes (Aggleton, 1985); in the others, small three-dimensional objects were used as stimulus material (Mumby, Pinel, & Wood, 1990; Rothblat & Hayes, 1987). In two experiments, monkeys' visual or tactual memory for objects was studied in the Wisconsin General Test Apparatus (Moss et al., 1988; Murray & Mishkin, 1984). Many similar examples are available, but these two were chosen because individual scores were tabulated from slightly larger samples of monkeys than is typical, because both normal and "pathological" (e.g., brain-damaged) subjects were included, and because the experiments were carried out in different laboratories.

Table 1 is a summary of observed/expected variance ratios for sets of data points. I have grouped together data points that the authors present in single figures or tables, corresponding to several conditions (e.g., serial positions, retention intervals) within the same experiment. However, data from distinct groups of animals, or from the same group tested before and after a brain operation, are in separate rows. For each set of data points, I show the number of trials per data point per subject (with T as defined above), the number of data points in the set, and summaries of variance ratios and chi-square statistics.

For example, the first row of Table 1 shows, from Murray and Mishkin's (1984) Table 1, results from a group of 4 rhesus monkeys prior to hippocampal ablation ("preop."). Their memory for single objects was tested in 100 trials at each of four retention intervals; hence there were four data points. To illustrate the calculation of observed and expected variance: at the 30-sec retention interval, the 4 monkeys made 94, 89, 93, and 95 correct responses, so the mean was 92.75/100 and the variance 6.92. Expected variance under the binomial model, given $T = 100$ trials and $P = .9275$, is $100 \times .9275 \times .0725 = 6.72$, and the observed/expected ratio is thus 1.03. Of course not all ratios are so close to 1; across all four data points, the ratios are 0.19, 0.35, 1.03, and 1.28. The table displays the median and range of these values.

To test whether the ratios deviate from the expected range around 1, each is converted to chi square by multiplying by three (number of subjects, N, minus 1) and

Table 1
Nonspatial Item Memory, Nonautomated Apparatus: Summary of Variability in Previous Studies

Source	Subjects	Trials per Data Point per Subject	No. Data Points	Observed/Expected Variance		No. χ^2		Pooled $\chi^2(df)$
				Median	Range	Low	High	
			Murray & Mishkin (1984)					
Table 1	4 monkeys, preop.	100	4	.69	.19-1.28	1		8.54(12)
	4 monkeys, postop.	100	4	1.02	.27-1.65			11.90(12)
	5 monkeys, preop.	100	4	1.64	1.21-1.91			25.64(16)
	5 monkeys, postop.	100	4	3.24	1.69-4.53		3	50.82(16)*
Table 2	3 monkeys, op. 1	100	12	1.74	0-6.53	1	4	50.24(24)*
	4 monkeys, op. 1	100	12	1.27	.67-4.07		3	60.69(36)*
			Aggleton (1985)					
Figure 2	6 rats	50	3	.58	.51- .61			8.53(15)
Figure 5	5 rats	50	13	.94	.10-2.78	1	1	48.95(52)
			Rothblat & Hayes (1987)					
Figure 2A	16 rats	12	10	1.07	.42-1.64	1		168.30(150)
Figure 2B	7 rats	30	2	.30, 0.66				5.77(12)
Figure 2B	5 rats	30	2	.73, 1.90				10.54(8)
			Moss, Rosene, & Peters (1988)					
Table 1	6 young monkeys	100	4	2.70	1.47-3.01		3	49.45(20)*
	6 older monkeys	100	4	2.12	1.09-3.89		2	46.15(20)*
			Mumby, Pinel, & Ward (1990)					
	14 rats	25	5	.87	.54-1.19			56.54(65)
			Reed, Chih-Ta, Aggleton, & Rawlins (1991)					
Figure 3	11 rats	6	5	.26	.11- .43	4		14.47(50)†
Figure 4	11 rats	6	10	.21	.12- .43	9		24.43(100)†

Note—$p > .95$ for number of χ^2 values that are too low and $p < .05$ for number of χ^2 values that are too high on the basis of expected variance. *Pooled χ^2 is significantly too high (values may be inexact). †Pooled χ^2 is significantly too low.

tested against chi-square tables with 3 *df*. The $p > .95$ and $p < .05$ critical values are 0.72 and 7.81, respectively. Under the null hypothesis that each set of scores is from a population having the specified expected variance, we would expect about 1 in 20 chi-square values to fall below 0.72, and about 1 in 20 to exceed 7.81. The smallest chi-square value is $0.19 \times 3 = 0.57$, which falls in the lower rejection region; none of the other 3 falls into either rejection region. This result is indicated in the table under the heading "No. χ^2 Low High," in this case 1 low and none high. The final column is a pooled chi-square obtained by summing chi-square and the degrees of freedom across all four data points; the result, 8.54 on 12 *df*, does not fall beyond either cutoff for 12 *df* (5.23, 21.0) meaning that the four variances collectively are consistent with the binomial model.

The distribution of individual chi-square values, and their pooled composite, can only be clearly interpreted if the tests being combined are independent; but here they represent repeated measures of variance on the same group of subjects, which are independent only if the deviations contributed by a particular subject to the separate variance measures are uncorrelated. The deviations will generally *not* be uncorrelated if genuine between-subject differences in mean performance exist, so that a particular subject tends to score consistently above or below the group mean. As outlined above, such between-subject differences will generate variances and chi-square values greater than the binomial expectation. However, where variances are close to or below the binomial expectation, there is no reason why the deviations that an individual contributes to the several measures of variance should be correlated (they represent random error only), so such sets of tests can be treated as independent. Composite chi-square values that fall in the upper and lower rejection regions are marked with asterisks and daggers, respectively.

In short, cases in which variances are markedly larger than expected from the binomial model are identified in the tables by a surplus of values in the "χ^2 High" column, and/or asterisks for the composite test; these results are descriptive rather than precise, because of possible nonindependence. The more controversial cases, in which variances are excessively low, have a surplus in the "χ^2 Low" column and/or a composite chi-square marked with daggers. If neither an asterisk nor a dagger appears, the binomial model satisfactorily accounts for that set of variances. (But note that the pooled tests are conservative if low variances are consistently associated with some, but not all, data points.)

Consider, first, the two studies of macaque monkeys' object recognition. In Murray and Mishkin (1984), the variance conforms to expectation for the two groups of animals tested prior to surgery. But of the four sets of

data that Murray and Mishkin obtained postoperatively (which came, of course, from the same animals as did the preoperative scores), three sets show evidence of larger than expected variance. For the group of 5 monkeys (row 4 of Table 1), 3/4 data points have variance ratios exceeding the $p < .05$ criterion; two other operated groups (rows 5 and 6) yield 4/12 and 3/12 ratios, which exceed it; and all the composite chi-square values are very large. This pattern of results supports the earlier suggestion that data from animals with brain lesions might be more variable than expected from the binomial model. The only operated group for which that is not so (row 2) was in fact unimpaired at the task after surgery.

That is not simply to make the point that brain-damaged animals are more variable than controls. The main implication is that the variability among *normal* animals is, in this case, adequately accounted for by the natural variance of binomially distributed scores. Individual variation in the underlying parameter P is not substantial. By contrast, the data of Moss et al. (1988), also from rhesus macaques tested at four retention intervals, give evidence of variance significantly larger than expected, among both normal (young adult) and older monkeys. As noted above, this is not at all surprising. There could be a number of reasons for the difference from Murray and Mishkin's results—for example, genuinely greater variance in memory ability in the subject population, or greater variation in procedural parameters.

The reports of experiments with rats for the most part present standard errors graphically. Having estimated from the graphs the standard errors of mean numbers of correct choices, I computed variances as $N \times (SE)^2$, where N is the number of subjects.

Aggleton (1985) studied 6 rats' nonmatching with single stimuli (complex goalboxes) at three retention intervals and, with a separate group of 5 rats, performance in 13 successive trial blocks of acquisition of matching. These were selected because they are the only portions of the data for which standard errors are displayed. As shown in Table 1, for neither data set does the variance deviate significantly from expectation (see also Rawlins et al., in press). Rothblat and Hayes (1987) and Mumby et al. (1990) tested rats' memory for single "junk" objects in a nonmatching paradigm; the data shown come from successive trial blocks of acquisition, and from later testing at various retention intervals. Again, the variances are consistent with the binomial model.

The scores reported by Reed et al. (1991), however, deviate extremely and consistently from expectation, as previously mentioned. The variance at every one of 15 data points is less than half the expected value, being significantly less in 13 cases. Some of the possible causes are discussed in the section below on Implications From Analysis of Variability (see also E. A. Gaffan & D. Gaffan, in press).

In summary, much of the published data on normal monkeys' and rats' memory for nonspatial items, from nonmatching or matching paradigms with manually operated apparatus, shows a degree of variability equal to, or somewhat larger than, what would be expected from the simple binomial model where all animals have a similar value of P. In only one case, Reed et al. (1991), is the reported variance significantly smaller than expected. This cannot be attributed to a calculation error, because the article presented individual rats' scores.

Nonspatial Item Memory, Automated Apparatus

In all of the experiments summarized in Table 2, highly practiced subjects were presented with lists of visual stimuli. The lists were initiated by the subject's making an observing response, but thereafter both stimulus presentation and response recording were automated.

W. A. Roberts and Kraemer (1981) gave squirrel monkeys and people three- and six-item lists of black-and-white patterns; 3 of the same monkeys were later tested with lists of three colored pictures (W. A. Roberts &

Table 2
Nonspatial Item Memory, Automated Apparatus: Summary of Variability in Previous Studies

Source	Subjects	Trials per Data Point per Subject	No. Data Points	Observed/Expected Variance		No. χ^2		Pooled $\chi^2(df)$
				Median	Range	Low	High	
			Roberts & Kraemer (1981)					
Figure 3	4 squirrel monkeys	384	9	1.90	.42- 5.03		4	60.54(18)*
Figure 5	4 people	384	9	11.55	1.42-23.6		8	289.90(18)*
			Roberts & Kraemer (1984)					
Figure 1	3 squirrel monkeys	144	12	1.49	.40- 3.20		1	38.43(24)*
Figure 2	3 squirrel monkeys	96	12	.89	.02- 6.35	1	3	44.52(24)*
Figure 3	3 squirrel monkeys	96	9	2.63	.19-13.85		4	80.56(18)*
Figure 4	3 squirrel monkeys	96	9	2.57	.30- 5.25		4	44.24(18)*
			Santiago & Wright (1984)					
	4 pigeons	40	24	1.80	.39- 4.77		5	129.3 (24)*
			Wright, Santiago, Sands, Kendrick, & Cook (1985)					
	6 people	20	8	3.98	2.07-7.56		6	168.5 (40)*

Note—$p > .95$ for number of χ^2 values that are too low and $p < .05$ for number of χ^2 values that are too high on the basis of expected variance. *Pooled χ^2 is significantly too high (values may be inexact).

Kraemer, 1984). Santiago and Wright (1984) gave pigeons lists of four colored pictures. (A similar procedure was used by Wright, Santiago, & Sands, 1984, with rhesus monkeys, but I omit it because there were only 2 subjects, causing chi-square tests to have extremely low power.) Wright et al. (1985) add comparable data from human subjects.

All the studies report performance at every serial position; for example, the 9 data points from W. A. Roberts and Kraemer (1981) collected in row 1 of Table 2 represent all positions of three- and six-item lists, and the 12 data points of W. A. Roberts and Kraemer (1984) are from three-item lists given under four different conditions of presentation. In general, variances have been estimated from graphs of the performance of individual animals, except in the case of the human data of Wright et al. (1985), who show only standard errors averaged across four serial positions for each of eight retention-interval conditions (thus there are only 8 "data points").

The analysis of observed/expected variance ratios for these experiments yields a consistent picture; in every case, variance is considerably greater than expected. (The tests are approximate for Wright et al., 1985, where the ratios were obtained by dividing the reported average variance across four serial positions by the average of the expected variances calculated for each serial position. However, the excess of observed over expected variance is so great that exactitude is unnecessary.)

It can be concluded that, for all species tested, variability at all or some serial positions is substantially greater than expected from the binomial model, the most likely cause being between-subject variation in the probability *P*. This is hardly surprising in view of the difficulty of these picture memory tasks, which necessitated lengthy training; Table 1, by contrast, implies that such large individual variance is not evident in all paradigms. The effect is not simply a consequence of having few subjects, as can again be ascertained through inspection of Table 1.

Spatial Item Memory, Nonautomated Apparatus

In all of the studies shown in Table 3, except that of Dale (1987), rats were tested in radial mazes. There have been many studies of this kind, but only the few in which standard errors have been reported can be considered here. The unpublished experiment by Caunt (1990) is included because the raw data are available.

Rats were allowed to enter four or five different arms of a maze, and then to choose between one of those and an as yet unentered arm. In Dale's (1987) experiments with college students, the subject sat in the center of a circular array of eight lamps. Seven of these were illuminated in sequence; then two, including the remaining lamp, were lit, and the task was to point to the one that had previously been illuminated. So, although stimulus presentation was automated, responses were recorded manually. Kesner and his colleagues (Kesner et al., 1988;

Table 3
Spatial Item Memory, Nonautomated Apparatus: Summary of Variability in Previous Studies

Source	Subjects	Trials per Data Point per Subject	No. Data Points	Observed/Expected Variance		No. χ^2		Pooled $\chi^2(df)$
				Median	Range	Low	High	
Bolhuis & van Kampen (1988)								
Experiment 1, Part 1	18 rats	5 or 10	6	1.19	1.05-1.64		1	127.89(102)*
Part 2	18 rats	2 or 4	9	1.16	.88-1.72		1	186.82(153)*
Part 3	18 rats	2	8	1.21	.88-2.07		3	194.36(136)*
Part 4	18 rats	2	8	1.06	.90-1.54			158.57(136)
Experiment 2	18 rats	3	8	1.13	.60-1.80		1	157.03(136)
Caunt (1990)								
	10 rats	6	3	.99	.85-1.11			26.66(27)
Kesner & Gray (1989)								
	5 rats, preop.	8	4	.85	.13- .90	1		10.92(16)
	5 rats, postop.	8	4	1.06	.68-1.80			18.40(16)
Kesner, Crutcher, & Beers (1988)								
Figure 1	7 rats	8	10	.98	.43-7.36		2	102.47(60)*
Figure 4	4 rats, preop.	8	5	1.62	.95-3.61		1	28.83(15)*
	4 rats, postop.	8	5	.40	.19-3.12		1	14.64(15)
Figure 5	4 rats, preop.	8	5	.63	.18- .92			8.71(15)
	4 rats, postop.	8	5	.39	.19-1.34			8.41(15)
Figure 7	4 rats, preop.	8	5	.93	.33-3.22		2	24.26(15)
	4 rats, postop.	8	5	1.00	.34-4.04		1	22.46(15)
Figure 8	6 rats, preop.	8	5	1.21	.89-2.49		2	39.53(25)*
	6 rats, postop.	8	5	1.59	.58-4.23		1	45.78(25)*
Dale (1987)								
Experiment 1	32 people	7	1	1.45				
Experiment 2	92 people	7	1	1.23				

Note—$p > .95$ for number of χ^2 values that are too low and $p < .05$ for number of χ^2 values that are too high on the basis of expected variance. *Pooled χ^2 is significantly too high (values may be inexact).

Kesner & Gray, 1989) and Dale used a stay contingency, where the familiar test item was the correct choice; Bolhuis and van Kampen (1988) and Caunt (1990) employed the opposite shift contingency, which rats find easier to learn.

As Table 3 shows, these experiments, including that on human subjects, generated data whose variance is near or slightly above that expected from the binomial model. The "observed" variances from Dale (1987) are in fact estimated means of within-cell variances obtained by calculation from reported F ratios and cell means, and they involve considerable approximation; so it would be difficult to test statistically the ratios between "observed" variance and the averaged expected variance across cells. It is interesting to note, however, that in both experiments, the resulting ratio is only slightly greater than 1.

Bolhuis and van Kampen (1988) trained the same 18 rats over two successive experiments, giving free choices during exposure to the list in Experiment 1 and forced choices in Experiment 2. Variances were higher than expected early in training and seem to have decreased toward those expected from the model later, possibly reflecting the effect of practice or the change of procedure. Four out of eight subgroups in Kesner et al. (1988) showed variances greater than expected. That may reflect the fact that the task they used, five-item lists with a stay contingency, was probably the most difficult—Kesner and Gray (1989) also used a stay contingency, but the list length was only four. Surprisingly, in contrast with the findings of Murray and Mishkin (1984, Table 1), the excessive variance is equally evident in operated and unoperated rats. This is not because the lesions were without effect, since all produced severe impairments in performance.

The overall conclusion is that the studies in this group did not yield any evidence of improbably low variance. As in the experiments on nonspatial item memory reviewed in Table 1, performance is reasonably well approximated by the binomial model, though, uncontroversially, there is evidence of additional variance in some cases.

Spatial Order Memory

Five studies of memory for order of spatial events, which allow estimation of observed/expected variance ratios, are presented in Table 4. They are rather diverse. In three, Kesner and his colleagues (Kesner & Gray, 1989; Kesner et al., 1984; Kesner & Novak, 1982) tested rats' memory for lists of radial maze arms; Dale (1987) gave college students lists of seven stimuli in the eight-lamp apparatus described above. The general procedure for testing order memory has been described in the section on determinants of serial position effects. Shimp (1976) gave pigeons a rather different task, which is included here because it required memory for the order in which spatially separated lamps had been illuminated in a three-item sequence. The last study was the only one in which subjects' responses were recorded automatically.

Shimp's (1976) data need no detailed consideration; the variance calculations are approximate because the numbers of trials per data point can only be estimated from the information provided in the paper. However, as in the other studies in which small numbers of subjects received lengthy training in automated apparatus (Table 2), variance was often greater than expected from the binomial model. The observed/expected ratio from Dale (1987), which is also approximate for the reasons given

Table 4
Spatial Order Memory: Summary of Variability in Previous Studies

Source	Subjects	Trials per Data Point per Subject	No. Data Points	Observed/Expected Variance		No. χ^2		Pooled $\chi^2(df)$
				Median	Range	Low	High	
			Shimp (1976)					
Condition 1	3 pigeons	574	3	1.13	.30-1.24			5.33(6)
Condition 2	3 pigeons	326	3	.78	.66-2.95			8.77(6)
Condition 3	3 pigeons	116	3	1.29	.68-8.07		1	20.06(6)*
Condition 4	3 pigeons	108	3	.60	.40-6.18		1	14.35(6)*
Condition 5	3 pigeons	67	3	.86	.68-5.6		1	14.25(6)*
Condition 6	3 pigeons	332	3	.10	.02-2.72			5.68(6)
			Kesner & Novak (1982)					
	4 rats, preop.	12 or 5	6	.61	.01-1.24	2		10.53(18)
	4 rats, postop.	12 or 5	6	.25	.05- .53	2		4.47(18)†
			Kesner, Measom, Forsman, & Holbrook (1984)					
	7 rats	8	3	.25	.13-1.01	2		8.36(18)†
			Kesner & Gray (1989)					
	5 rats, preop.	8	3	.33	.30- .76			5.56(12)
	5 rats, postop.	8	3	.97	.96-1.60			14.12(12)
			Dale (1987)					
Experiment 2, Table 4	93 people	7	1	1.65				

Note—$p > .95$ for number of χ^2 values that are too low and $p < .05$ for number of χ^2 values that are too high on the basis of expected variance. *Pooled χ^2 is significantly too high (values may be inexact). †Pooled χ^2 is significantly too low.

above, is slightly larger than those obtained from his experiments on item memory (Table 3) but nonetheless not much greater than 1.

The data of Kesner and Novak (1982) on rats' memory for eight-item lists, and of Kesner et al. (1984) and Kesner and Gray (1989) for four-item lists, are problematic. There is a clear tendency for variance (as estimated from graphed standard errors) to be less than expected from the binomial model. Of 21 data points across the three studies, 12 have variance less than half the expected value, and in 6 cases the variance is significantly below expectation (compared with about 1 case expected by chance). Two subgroups yield composite chi-square values smaller than the .95 significance criterion; Kesner and Gray's preoperative data are marginal, but probably within the range of measurement error in my estimates from graphs.

The discrepancies are not so extreme as those seen in the case of Reed et al. (1991; cf. the present Table 1), and of course it is impossible to comment on the three other published papers from Kesner's laboratory (Kesner et al., 1987; Kesner et al., 1986; Kesner & Holbrook, 1987), in which similar U-shaped serial position curves are presented without standard errors. However, taken in conjunction with a reported failure to replicate Kesner and Novak's (1982) results (see Maki et al., 1984), the analysis suggests that these serial position effects should be treated with caution. Some possible explanations for the unexpectedly low variance are discussed below.

IMPLICATIONS FROM ANALYSIS OF VARIABILITY

The first conclusion is that the simple model, predicting that numbers of correct responses by a group of subjects should be binomially distributed in accordance with a single value of P corresponding to the group's mean score, approximates reasonably well the results of many studies of spatial and nonspatial memory in which mazes or other manually operated apparatuses have been used (Tables 1, 3, and 4; for further examples, see Rawlins et al., in press). These are tasks for which subjects require relatively little training. However, it is fairly common for variance to be greater than expected from the model. The most likely—and unsurprising—reason is that the subjects differ in their true values of P. There is inconsistent support for the speculation that such individual differences would be more prominent in groups with brain lesions (Tables 1, 3, 4), but more support for the idea that they would appear when the task requires long training in an automated apparatus (Tables 2 and 4). These general conclusions apply regardless of whether rats, pigeons, monkeys, or people are the experimental subjects.

Various recommendations follow from these findings, regarding the numbers of trials and/or subjects that are, or should be, used in experiments of this kind (see Conclusions, below).

As for the relevance of the analysis to the evidence for serial position effects, much of that evidence stands up well to statistical scrutiny. The reports of both primacy and recency in tests of visual item memory in automated apparatuses (Table 2) or of spatial item memory in radial mazes (Table 3) show degrees of variability that are in accordance with the statistical assumptions I have made. Conformity with the binomial model is not in itself a criterion of validity—but, as will be shown below, several threats to validity can in principle be detected, because they give rise to excessively low variability.

Two of the claims reviewed above have been shown to be questionable in that regard. The demonstrations of primacy as well as recency in rats' memory for lists of nonspatial items (Reed et al., 1991) and for the order of items in a spatial list (Kesner & Novak, 1982, and other studies shown in Table 4) depend on data whose variability (when it has been possible to assess it) is pervasively smaller than expected from the binomial model.

I will argue below that two possible causes of low variance are nonindependence of trials and restricted selection of data; that there is evidence for these states of affairs among the studies that have evinced low variance; and that the validity of those experiments is suspect as a consequence.

Before I discuss these problems, I must point out that there is one possible circumstance that may reduce variance below the binomial expectation, but that does not necessarily invalidate an experiment. Animals, rather than each one having an individual value of P that is constant across trials as the binomial model assumes, may switch between "states" so that the value of P is high on some trials and low on others. An extreme example would be a rigid position habit, whereby an animal always chooses correctly ($P_1 = 1.0$) when the correct alternative is on its preferred side, and incorrectly ($P_2 = 0$) when it is on the opposite side. Another hypothetical case would occur when an animal discriminates well on some trials (say, $P_1 = .9$) but chooses randomly ($P_2 = .5$) on others.

Suppose that the numbers of trials of the two types are T_1 and T_2, and that these and also P_1 and P_2 are similar across animals. An animal's total score will be the sum of two binomial variables with probabilities P_1 and P_2, and the variance of the total score the sum of their two variances—that is, $T_1P_1(1-P_1)+T_2P_2(1-P_2)$. It can be shown that this will generally be smaller than the expectation from our model, $(T_1+T_2)P(1-P)$, where P is the average resulting from the mix of trials with P_1 and P_2. The reduction will be steep when P_1 and P_2 are widely different, with the average near chance, as in the "position habit" example, but rather slight in the second type of case; when $P_1 = .9$, $P_2 = .5$ and $T_1 = T_2$, variance will be about 0.8 of that expected from our binomial model. Average values of P in memory experiments are mostly above chance, so the second example would be more representative. Inspection of Tables 1 and 4 shows that the observed/expected ratios in the studies by Reed et al. (1991) and by Kesner and his colleagues (Kesner

& Gray, 1989; Kesner et al., 1984; Kesner & Novak, 1982) are generally much smaller than 0.8. So it is necessary to consider other explanations.

Nonindependence of Trials

The probability that a subject responds correctly on a given trial is determined by systematic effects (of the serial position, the ability of the subject, etc.) and error. The error component is assumed to vary randomly and independently across trials; binomial variance is one source of error. If trials are nonindependent, this means that the error component is to some extent predictable, either from the response of another subject on the corresponding trial (nonindependence between subjects) or from the response of the same subject on a preceding trial (nonindependence within subjects). In either case, to the extent that the error component is not free to vary randomly across trials, error variance will be less than it should be.

Nonindependence between subjects could arise if, for example, the assignment of stimuli to conditions is not randomized across subjects, but exactly the same sequences of stimuli are used in corresponding lists and choice tests for every subject. That was apparently the case in Reed et al.'s (1991) study (see p. 39), but one cannot ascertain from published accounts whether it applied to the experiments of Kesner and his colleagues. If all subjects have strong and similar preferences between stimuli, their choices on corresponding trials will be correlated, and error variance will be reduced. Inspection of the data trial by trial will show whether subjects' choices are correlated. If so, a possible artifact exists, because when stimulus assignment is completely confounded with conditions (serial positions), it is possible that observed differences between serial positions are coincidentally generated at least in part by stimulus preferences. Such coincidences are unlikely when many different stimuli are tested at each serial position, but when as few as five to eight trials are given per serial position, as is often the case in maze experiments (see Tables 1, 3, and 4), it is a potential flaw that should be eliminated by properly randomizing stimulus assignment. Another possible source of nonindependence between animals is experimenter expectation, which could result in the assignment of unduly similar scores to different animals.

In fact, inspection of Reed et al.'s (1991) raw data has not yielded evidence of between-animal correlation, but of the alternative form of nonindependence—namely, within-animal correlation (Rawlins et al., in press). Subjects' performance in a certain condition was negatively correlated with their own score cumulated from earlier sessions. Such a state of affairs is unlikely to be generated by animals themselves, but it could result from experimenter expectation or bias, producing overall scores that converge on an expected total.

Kesner and his colleagues' data have not been analyzed trial by trial, so it is unknown whether they show any of the forms of nonindependence described above. However, there are clear indications that a different factor could have been responsible for restricting variance in their studies—selection of data.

Selection of Data Through Training Criteria

In most of the experiments by Kesner and colleagues on rats' spatial memory, a performance criterion was imposed during initial training, which would probably have the effect of reducing variance below its expected value at some data points. Such a criterion is unacceptable for several reasons.

In the studies of both item and order memory by Kesner and his colleagues (Kesner et al., 1987; Kesner et al., 1989; Kesner et al., 1988; Kesner et al., 1986; Kesner & Gray; 1989; Kesner & Holbrook, 1987; Kesner et al., 1984), rats were initially trained until they had scored 75% correct across the most recent eight tests of both the first and the last list positions. At most, 100 training trials were given. In some papers, it is stated that a rat's data were excluded from analysis if the rat did not reach criterion; in others, this is not explicitly stated but we may infer that it was so. In two studies (Kesner et al., 1986; Kesner et al., 1984), the number of rats that failed to reach criterion is reported (a total of 10 out of 66 trained alike on eight-arm lists), but in other studies in which the same task was given, there is no mention of rats' being dropped from the experiment. With the easier four- or five-arm versions of the tasks, it is sometimes stated (e.g., by Kesner et al., 1984) that all rats reached criterion within 100 trials, but that is not always made clear.

The criterion of 75% correct at both first and last serial positions could, of course, be attained by animals that performed well at all positions, or that showed both primacy and recency, but not, for example, by animals that showed recency alone. If there are such animals, they are represented scantily or not at all in the group statistics.

The practice of presenting animals' data only from the trial block in which they reached criterion, and of not testing them past that point (except postoperatively), implies that the group mean score must, trivially, be at least 75% correct at both the first and the last serial positions, though it could be lower at intermediate positions. Worse, the procedure capitalizes on chance. The 75% criterion was based on serial samples of eight trials, so by chance fluctuation a rat might reach "criterion" within that short run of trials and contribute to the picture of "primacy" and "recency" in the group means. If such a rat were tested longer, it would not necessarily maintain criterial performance.

The argument of Kesner et al. (1984, p. 379), that the observed mean of trials to criterion was too small to be attained by chance alone, is unconvincing, because it is based on an arbitrary value, .5, for the probability of reaching criterion by chance on any one trial. In the same paper, Kesner et al. address this problem (p. 380) by arguing that sham- or unoperated animals maintain the previous pattern of performance when retested after reaching criterion; but in cases of this kind that are mentioned in published papers, there is no statistical confirmation

of serial position effects (Kesner et al., 1988; Kesner et al., 1986; Kesner & Holbrook, 1987; Kesner & Novak, 1982). It should be routine procedure to present and analyze postcriterion data from all subjects, not only data from the "criterion run" itself.

Selectivity among subjects might perhaps be justified in neuropsychological terms, if the aim were to test the effects of a lesion on a psychological function (e.g., primacy) only in subjects that manifest the function in the first place. However, the argument that the claimed primacy and recency effects are affected by chance fluctuation devalues that rationale and opens up the possibility that at least part of the postoperative reduction in memory for beginnings and ends of lists reflects simple regression to the mean.

It is obvious that the training criterion might result in unduly low variance at the first and last serial positions, where scores were constrained between 75% and 100% correct. If that is the primary cause, the observed/expected variance ratios should be smallest at the serial positions to which the criterion applied, the first and last during preoperative training only (postoperative testing was based on a fixed number of trials with no criterion). Inspection of the original data partly supports that explanation. In the preoperative results of Kesner and colleagues (Kesner et al., 1988; Kesner & Gray, 1989; Kesner et al., 1984), the ratios are somewhat lower at the first and last positions than they are elsewhere. But the lowest ratios of all are seen in both the pre- and the postoperative data of Kesner and Novak (1982), who trained all subjects for the same number of trials, mentioning no criterion. None of the procedural details in the report suggests an obvious explanation for the low variance, so it would be pointless to speculate further.

Because the above-mentioned training criterion per se would restrict variance only at some serial positions, it should not necessarily result in significant deviations when data are pooled across all serial positions. Thus, although I have used low variance as an indicator of possible problems in some of the studies done by Kesner and colleagues, it is the procedure itself that is primarily objected to, whether it generates consistently low variance (in some order memory studies; see Table 4) or not (in the item memory studies; see Table 3). The fact that low values of pooled chi-square are evident mainly in spatial order, not spatial item memory, may indicate simply that the item memory data typically include more serial positions, so a smaller proportion of the data being pooled suffers from restricted variance. Alternatively, the serial position differences in item memory may be more genuine, and thus would emerge even without data selection (the primacy effects found by Bolhuis & van Kampen, 1988, and replicated by Caunt, 1990, appeared in an item memory procedure, though under different conditions from Kesner's). However, it must be reiterated that data influenced by the training criterion are of no value for establishing the true magnitude of primacy and/or recency effects in the normal population of rats. To the extent that such effects do exist, Kesner's methodology must exaggerate them.

CONCLUSIONS AND RECOMMENDATIONS

Diagnosing Experimental Faults

I have argued that a variety of problems, such as experimenter effects and biases in data selection, can give rise to lower than expected variance in scores that should be binomially distributed. One may recommend, therefore, that researchers collecting such data compare the observed variance of scores with that expected under the binomial model, especially when—as is generally the case with tests of serial position effects in manually operated apparatuses—testers are not blind to conditions and/or experimental hypotheses.

Such comparisons are of course only worthwhile if the additional contribution of between-subject variance is small; otherwise it would be difficult to detect any reduction resulting from nonindependence of trials, and so forth. As I have shown above, between-subject variance is substantial in automated experiments, but small and sometimes negligible in typical manual apparatuses such as mazes. It is when experiments are not automated that the problems giving rise to reduced variance are most likely to occur.

Statistical Power

From Tables 1, 3, and 4, one can see that, in maze experiments, serial position curves have often been constructed on the basis of rather few trials per subject per data point. It is instructive to consider how many trials are needed to give a reasonable chance of detecting differences of the magnitude typically seen within serial position curves.

Suppose, first, that there is only 1 subject, and that we wish to detect a primacy effect that reflects true probabilities of .7 and .5 of making a correct choice, at early and middle list positions, respectively. On the reasonable assumption that all trials are independent, within and between serial positions, we can use statistical power tables for binomial proportions (Cohen, 1988, chap. 6) to judge how many trials would be required to reach the conventionally acceptable 80% chance of detecting a significant difference between proportions of .7 and .5. If we were willing to assume that the *opposite* of a primacy effect was unlikely across those serial positions—that is, that it was implausible for the "early" probability to be truly less than the "middle" one—we could carry out a one-tailed test. Setting $\alpha = .05$, about 80 trials per serial position would be required for a one-tailed test, and about 100 trials for a two-tailed. This recommendation is, of course, only a plausible illustration; Cohen (1988) shows the necessary adjustments for different values of P, and so forth.

These figures indicate what is needed to compare serial positions reliably in single subjects, as is often attempted with automated apparatus. Table 2 shows that W. A. Roberts and Kraemer (1981, 1984) gave adequate numbers of trials per serial position, whereas Wright and colleagues (Santiago & Wright, 1984; Wright et al., 1985) gave rather few. That problem is mitigated because Wright

and colleagues display several serial position curves per subject and are generally concerned with the change in shape across curves, not specific comparisons within curves. However, the detailed discussion of the form of serial position functions of individual monkeys by Wright et al. (1984, p. 523) is not justified, because only 20 trials were run at each position, so the experiment had low power to detect such differences reliably.

In most experiments done with rats, groups, not individuals, are tested, but the binomial model allows us to estimate the numbers of trials required. My analysis suggests that much data from normal rats in mazes conforms reasonably with the model's assumptions of common values of P across animals and independence of trials between and within animals. If so, 10 trials from each of 8 animals (for example) are equivalent to 80 trials from a single animal, for the purpose of estimating or comparing values of P. The total of trials across all subjects (number of subjects $\times$ number of trials per serial position, $N \times T$) should therefore equal the required number above, approximately 80 trials for a one-tailed test in my illustration.

The maze experiments in Table 3 (spatial item memory) that were done with rats and that did conform to the model had fewer subjects and trials than the rough criterion suggested. The experiments on rats' spatial order memory in Table 4 also had too few subjects and trials for adequate power; but I have argued above that the variance was smaller than expected from the binomial model, so the power calculations do not apply. If an experiment in reality has low power to detect differences among conditions, a procedure that artificially restricts variance will increase the likelihood of obtaining "significant" differences, but that apparent increase in power is of course illusory.

In many experiments shown in the tables, the variance was greater than was expected from the binomial model. If, as is most likely, this stems from additional between-subject variance in P, its effect on power is difficult to estimate precisely. When scores at different serial positions are compared in t tests or analyses of variances, part of that variance will contribute to the estimate of the overall between-subject effect, which is ignored when serial positions are compared within subjects. However, the increased variance within each serial position will also inflate the estimated subjects $\times$ serial position error variance to an unknown degree, so it would be prudent to adopt a conservative standard for the number of trials nonetheless.

Another strategy for increasing the power of tests of serial position effects despite relatively few trials is to conduct omnibus tests—as, for example, tests of quadratic trend across positions to verify the apparent curvature of a function. Of course, such tests, especially when they pool error variance across several serial positions, do not provide a satisfactory basis for testing differences between any particular pair of positions (O'Brien & Kaiser, 1985).

Evidence for Primacy and Recency

The foregoing discussion implies that much of the research that appears to demonstrate primacy in animal memory is open to serious doubt. By contrast, there is little reason for scepticism about recency, not only because the effect can be replicated easily, but because it is entirely consistent with other findings such as the well-established effect of a time lapse on recall of a single item.

Many reports of primacy are flawed either by excessively low variance or by unacceptable data selection policies. The only ones that are not vulnerable on these grounds are Bolhuis and van Kampen's (1988) study of spatial memory (see Table 3—specifically, Experiment 2, in which a forced-choice procedure was used) and the automated experiments on visual memory by W. A. Roberts and Kraemer (1981) and Wright and colleagues (Sands & Wright, 1980, and the subsequent papers shown in Table 2). However, as previously noted, in all of the automated experiments, an observing response was employed for the initiation of the list, so that these experiments are subject to D. Gaffan's objection (1983) that primacy reflects attentional rather than mnemonic enhancement.

No observing response is employed in experiments with rats in mazes, but only one such report of primacy (Bolhuis & van Kampen, 1988) demonstrably escapes criticism. Furthermore, such experiments frequently have inadequate statistical power, so there must be some doubt about the strength of the effect even when it is validly measured. Two published replications of the same procedure—memory for five-item lists in a stay contingency—were appropriately designed, because equal numbers of trials were given to all subjects and no data were discarded (DiMattia & Kesner, 1984, 1988). Primacy was observed in the former study, but not in the second.

If primacy in animals is of theoretical interest, I have tried to indicate what kinds of procedures are required to establish whether it exists, and to assess its magnitude under various conditions. The methodological and statistical principles that I have discussed can be applied more broadly to experiments on animal learning and memory in which choice among alternatives is used as the measure of performance.

REFERENCES

AGGLETON, J. P. (1985). One-trial object recognition by rats. *Quarterly Journal of Experimental Psychology*, **37B**, 279-294.

BOLHUIS, J. J., & VAN KAMPEN, H. S. (1988). Serial position curves in spatial memory of rats: Primacy and recency effects. *Quarterly Journal of Experimental Psychology*, **40B**, 135-149.

BUCHANAN, J. P., GILL, T. V., & BRAGGIO, J. T. (1981). Serial position and clustering effects in a chimpanzee's "free recall." *Memory & Cognition*, **9**, 651-660.

CAUNT, J. (1990). *Spatial memory in rats: Primacy and recency effects in the serial position curve.* Unpublished bachelor's thesis, University of Reading.

COHEN, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Erlbaum.

COOK, R. G., BROWN, M. F., & RILEY, D. A. (1985). Flexible memory processing by rats: Use of prospective and retrospective infor-

mation in the radial maze. *Journal of Experimental Psychology: Animal Behavior Processes*, **11**, 453-469.

DALE, R. H. I. (1987). Similarities between human and animal spatial memory: Item and order information. *Animal Learning and Behavior*, **15**, 293-300.

DIMATTIA, B. V., & KESNER, R. P. (1984). Serial position curves in rats: Automatic vs. effortful information processing. *Journal of Experimental Psychology: Animal Behavior Processes*, **10**, 557-563.

DIMATTIA, B. V., & KESNER, R. P. (1988). Role of the posterior parietal association cortex in the processing of spatial event information. *Behavioral Neuroscience*, **102**, 397-403.

GAFFAN, D. (1977). Recognition memory after short retention intervals in fornix-transected monkeys. *Quarterly Journal of Experimental Psychology*, **29**, 577-588.

GAFFAN, D. (1983). A comment on primacy effects in monkeys' memory for lists. *Animal Learning & Behavior*, **11**, 144-145.

GAFFAN, D., & WEISKRANTZ, L. (1980). Recency effects and lesion effects in delayed nonmatching to randomly baited samples by monkeys. *Brain Research*, **196**, 373-386.

GAFFAN, E. A., & GAFFAN, D. (in press). Less than expected variability in evidence for primacy and von Restorff effects in rats' nonspatial memory. *Journal of Experimental Psychology: Animal Behavior Processes*.

HAYS, W. L. (1988). *Statistics* (4th ed.). New York: Holt, Rinehart & Winston.

HITCH, G. J. (1983). Short-term memory processes in humans and animals. In A. Mayes (Ed.), *Memory in animals and humans: Some comparisons and their theoretical implications* (pp. 177-202). Wokingham, U.K.: Van Nostrand Reinhold.

HITCH, G. J. (1985). Short-term memory and information processing in humans and animals: Towards an integrative framework. In L.-G. Nilsson & T. Archer (Eds.), *Perspectives on learning and memory* (pp. 119-136). Hillsdale, NJ: Erlbaum.

KESNER, R. P., ADELSTEIN, T., & CRUTCHER, K. A. (1987). Rats with nucleus basalis magnocellularis lesions mimic mnemonic symptomatology observed in patients with dementia of Alzheimer's type. *Behavioral Neuroscience*, **101**, 451-456.

KESNER, R. P., ADELSTEIN, T., & CRUTCHER, K. A. (1989). Equivalent spatial location memory deficits in rats with medial septum or hippocampal formation lesions and patients with dementia of the Alzheimer's type. *Brain & Cognition*, **9**, 289-300.

KESNER, R. P., CRUTCHER, K., & BEERS, D. R. (1988). Serial position curves for item (spatial location) information: Role of the dorsal hippocampal formation and medial septum. *Brain Research*, **454**, 219-226.

KESNER, R. P., CRUTCHER, K. A., & MEASOM, M. O. (1986). Medial septal and nucleus basalis magnocellularis lesions produce order memory deficits in rats which mimic symptomatology of Alzheimer's disease. *Neurobiology of Aging*, **7**, 287-295.

KESNER, R. P., & GRAY, M. L. (1989). Dissociation of order and item memory following parietal cortex lesions in the rat. *Behavioral Neuroscience*, **103**, 907-910.

KESNER, R. P., & HOLBROOK, T. (1987). Dissociation of item and order spatial memory in rats following medial prefrontal cortex lesions. *Neuropsychologia*, **25**, 653-664.

KESNER, R. P., MEASOM, M. O., FORSMAN, S. L., & HOLBROOK, T.H. (1984). Serial-position curves in rats: Order memory for episodic spatial events. *Animal Learning & Behavior*, **12**, 378-382.

KESNER, R. P., & NOVAK, J. M. (1982). Serial position curve in rats: Role of the dorsal hippocampus. *Science*, **218**, 173-175.

MACPHAIL, E. M. (1980). Short-term visual recognition memory in pigeons. *Quarterly Journal of Experimental Psychology*, **32**, 521-538.

MAKI, W. S., BEATTY, W. W., & CLOUSE, B. (1984). Item and order information in spatial memory. *Journal of Experimental Psychology: Animal Behavior Processes*, **10**, 437-452.

MOSS, M. B., ROSENE, D. L., & PETERS, A. (1988). Effects of aging on visual recognition memory in the rhesus monkey. *Neurobiology of Aging*, **9**, 495-502.

MUMBY, D. G., PINEL, J. P. J., & WOOD, E. R. (1990). Nonrecurring-items delayed nonmatching-to-sample in rats: A new paradigm for testing nonspatial working memory. *Psychobiology*, **18**, 321-326.

MURRAY, E. A., & MISHKIN, M. (1984). Severe tactual as well as visual memory deficits follow combined removal of the amygdala and hippocampus in monkeys. *Journal of Neuroscience*, **4**, 2565-2580.

O'BRIEN, R. G., & KAISER, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, **97**, 316-333.

OVERMAN, W. H., MCLAIN, C., ORMSBY, G. E., & BROOKS, V. (1983). Visual recognition memory in squirrel monkeys. *Animal Learning & Behavior*, **11**, 483-488.

RAWLINS, J. N. P., DEACON, R. M., CHIH-TA, T., & AGGLETON, J. P. (in press). Doubts concerning primacy in rats' nonspatial recognition memory: A reply to Gaffan and Gaffan. *Journal of Experimental Psychology: Animal Behavior Processes*.

REED, P. (in press). Fewer doubts concerning rats' serial position performance: A reply to Gaffan and Gaffan, and Rawlins, Deacon, Chih-Ta and Aggleton. *Journal of Experimental Psychology: Animal Behavior Processes*.

REED, P., CHIH-TA, T., AGGLETON, J. P., & RAWLINS, J. N. P. (1991). Primacy, recency and the von Restorff effect in rats' nonspatial recognition memory. *Journal of Experimental Psychology: Animal Behavior Processes*, **17**, 36-44.

ROBERTS, S. (1987). Less-than-expected variability in evidence for three stages in memory formation. *Behavioral Neuroscience*, **101**, 120-125.

ROBERTS, W. A., & KRAEMER, P. J. (1981). Recognition memory for lists of visual stimuli in monkeys and humans. *Animal Learning & Behavior*, **9**, 587-594.

ROBERTS, W. A., & KRAEMER, P. J. (1984). Picture memory in monkeys. *Canadian Journal of Psychology*, **38**, 218-236.

ROBERTS, W. A., & SMYTHE, W. E. (1979). Memory for lists of spatial events in the rat. *Learning & Motivation*, **10**, 313-336.

ROTHBLAT, L. A., & HAYES, L. L. (1987). Short-term object recognition memory in the rat: Nonmatching with trial-unique junk stimuli. *Behavioral Neuroscience*, **101**, 587-590.

SANDS, S. R., & WRIGHT, A. A. (1980). Primate memory: Recognition of serial list items by a rhesus monkey. *Science*, **209**, 938-939.

SANTIAGO, H. C., & WRIGHT, A. A. (1984). Pigeon memory: Same-different concept learning, serial probe recognition acquisition, and probe delay effects on the serial position function. *Journal of Experimental Psychology: Animal Behavior Processes*, **10**, 498-512.

SHIMP, C. P. (1976). Short-term memory in the pigeon: Relative recency. *Journal of the Experimental Analysis of Behavior*, **25**, 55-61.

THOMPSON, R. K. R., & HERMAN, L. M. (1977). Memory for lists of sounds by the bottle-nosed dolphin: Convergence of memory processes with humans? *Science*, **195**, 501-503.

WRIGHT, A. A. (1989). Memory processing by pigeons, monkeys and people. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 24, pp. 25-70). New York: Academic Press.

WRIGHT, A. A., SANTIAGO, H. C., & SANDS, S. F. (1984). Monkey memory: Same-different concept learning, serial probe acquisition and probe delay effects. *Journal of Experimental Psychology: Animal Behavior Processes*, **10**, 513-529.

WRIGHT, A. A., SANTIAGO, H. C., SANDS, S. F., KENDRICK, D. F., & COOK, R. G. (1985). Memory processing of serial lists by pigeons, monkeys and people. *Science*, **229**, 287-289.

WRIGHT, A. A., & WATKINS, M. J. (1987). Animal learning and memory and their relation to human learning and memory. *Learning & Motivation*, **18**, 131-146.

(Manuscript received July 2, 1991;
revision accepted for publication March 16, 1992.)