Behavioral Neuroscience
1987, Vol. 101, No. 1, 120-125

0735-7044/87/$00.75

Less-Than-Expected Variability in Evidence for Three Stages in Memory Formation

Seth Roberts
University of California, Berkeley

Gibbs and Ng (1976, 1977) proposed a three-stage model of long-term memory formation on the basis of results from a one-trial-learning passive-avoidance procedure using chicks. Chi-square tests show that much of the evidence for this model (Gibbs & Ng, 1976, 1977, 1979, 1984a, 1984b), involving transient drops in retention and drug effects, is less variable than would be expected by chance. This should reduce belief in the model.

Lee-Teng and her colleagues (Cherkin & Lee-Teng, 1965; Lee-Teng & Sherman, 1966) developed a widely used one-trial-learning procedure involving chicks. The training trial consists of placing a small bead coated with a bitter-tasting substance close to the chick's beak; the chick usually pecks the bead. The test trial consists of a re-presentation of the bead. Memory for the training trial is shown by a reduced tendency to peck the bead on the test trial.

On the basis of work with this procedure, Gibbs and Ng (1976, 1977) proposed a three-stage model of memory formation. According to their model, the first stage of memory (short-term) is formed soon after the training trial and begins to decay after about 10 minutes. The second stage of memory (intermediate-term) is based on the first stage; it begins roughly when the first stage ends, and it lasts until about 50 minutes after the training trial. The final stage (long-term) begins roughly when the second stage ends, and it lasts indefinitely. Different stages are sensitive to different drugs; for example, ouabain acts on the second stage but not the first.

The main support for these ideas has come from two lines of evidence. (a) *Drug effects*: Gibbs and Ng (1976, 1977) measured the effect of drugs that reduced retention using a range of train-test intervals. (The train-test interval [TTI] is the time between the training trial and the testing trial.) Some drugs reduced retention only with TTIs greater than a certain time (e.g., 10 min). Categorized according to the necessary TTI, the drugs fell into three different groups, suggesting that they were acting on three different memories. (b) *Transient drops in retention*: In some experiments (Gibbs & Ng, 1979, 1984a, 1984b), the function showing retention as a function of TTI contains what look like fissures (e.g., Figure 1). All of the functions contain two fissures, which supposedly correspond to change-overs from short-term to intermediate-term memory (the left-hand fissure) and from intermediate-term to long-term memory (the right-hand fissure). Retention is low for a short period because one memory decays before the next becomes accessible.

The purpose of this article is to point out that much of this evidence is less variable than would be expected by chance, assuming that trials with different chicks are independent. The two kinds of evidence—transient drops in retention and drug effects—will be covered in separate sections.

Transient Drops in Retention

According to the procedure section for the data shown in Figure 1 (Figure 3 of Gibbs & Ng, 1984b),

> Day-old [chicks] were pretrained in pairs to peck at a red and a blue glass bead. . . . Following pretraining, a similar red bead . . . was dipped in the chemical aversant, methyl anthranilate, and presented for 10 s. . . . Tests for retention of the learned association were carried out at various times between 5 min and 24 h after learning, with a different group of approximately 20 chicks for each learning-retention interval under each drug condition. In these retention trials, a dry red and a dry blue bead were presented in succession for 10 s each. . . . The proportion of chicks avoiding [i.e., not pecking] the red bead at each retention trial was taken as an index of memory. (Gibbs & Ng, 1984b, p. 110)

Figure 1. Retention as a function of time between training trial and test trial for the bead used in training (red) and another bead (blue). (Corticosterone was given to all chicks shortly after the training trials.) *Note.* From "Hormonal Influence on the Duration of Short-Term and Intermediate Stages of Memory" by M. E. Gibbs and K. T. Ng, 1984, *Behavioural Brain Research, 11*, Figure 3, p. 112. Copyright 1984 by Elsevier Science Publishers. Reprinted by permission.

This work was supported by National Institute of Mental Health Grant 1 RO1 MH383358-01.

I thank Juliet Shaffer for statistical advice and Marc Breedlove and Mark Rosenzweig for comments on the manuscript.

Correspondence concerning this article should be addressed to Seth Roberts, Department of Psychology, University of California, Berkeley, California 94720.

Thus each group of chicks should contribute one red-bead point and one blue-bead point in Figure 1. At some TTIs (e.g., 15 min), the number of red-bead points (one) differs from the number of blue-bead points (two). This is because coincident points were not distinguished (K. T. Ng, personal communication, March 18, 1985). At a TTI of 15 min, for example, this means that two groups were tested and produced different results with the blue bead but the same results with the red bead. According to Ng, the lines summarizing the red-bead data were drawn in an unusual way—rather than through the average of the red-bead points—"to indicate more clearly the effects of replication and to show that replication was not carried out on all TTIs" (K. T. Ng, personal communication, March 18, 1985).

Three papers (Gibbs & Ng, 1979, 1984a, 1984b) contain a total of 11 graphs that resemble Figure 1. Two of the papers (Gibbs & Ng, 1979, 1984b) contain no other graphs; the other paper (Gibbs & Ng, 1984a) contains three other graphs that differ in many ways from Figure 1. If trials with different chicks were independent, then all of the graphs resembling Figure 1 show variability that is less than what is expected due to chance alone. When the dips in the function are excluded, the rest of the red-bead points are remarkably constant, and the blue-bead points are also unusually constant.

To estimate the probability that the data would be so constant by chance, let us assume that each point is based on exactly 20 chicks and that trials with different chicks were independent. Then chi-square tests allow us to compare the variability found in the data with the variability expected by chance. Consider the red-bead function shown in Figure 1. Including replications, it has 37 points. Based on comparisons with the number of blue-bead points, two of the red-bead points (at TTIs of 15 min and 70 min) represent two coincident points, increasing the total to 39 points. To eliminate the variability caused by the dips, I exclude the data from the bottom of the dips (at TTIs of 15 and 80 min, 2 points per TTI) and the points next to them (at TTIs of 10, 20, 75, and 85 min, 2 points per TTI), leaving 27 points. Each point corresponds to two counts—the number of chicks avoiding and the number not avoiding—so that the 27 points correspond to a 2×27 table of counts, where the 2 rows correspond to the numbers avoiding and not avoiding and the 27 columns correspond to the 27 different points. The null hypothesis is that the probability of an avoidance was the same for all 27 points. Because there are 26 degrees of freedom, the expected chi-square is 26. The actual chi-square is 10.4. If the null hypothesis is true, the probability of a value as low as this happening by chance is .004. If the null hypothesis is false—if the probability of an avoidance varies from point to point—the probability is even lower.

The blue-bead data can be analyzed in the same way, except that there is no need to exclude any of the points. The expected chi-square is 37; the actual chi-square is 20.4. The probability of a value as low as this happening by chance is .03.

Table 1 shows the result of analyzing all 11 graphs in this way. Two graphs (Figure 3 of Gibbs & Ng, 1979, and Figure 2 of Gibbs & Ng, 1984a) contain a dip that is at a minimum for two TTIs; in these two cases, the data from both TTIs were excluded, along with the data from the two neighboring TTIs. In two cases (Figure 1 of Gibbs & Ng, 1984a, and Figure 1 of Gibbs & Ng, 1984b), differences in the number of red-bead points and blue-bead points at a single TTI meant that there were some coincident points, but it was impossible to determine exactly what the coincident points were; in these cases, values were chosen that maximized the chi-square. Combining the analyses over graphs for each combination of bead (two beads) and paper (three papers) gives six overall

Table 1
Chi-Square Values for 22 Data Sets: Observed Versus Expected

	Red bead			Blue bead		
Source	Expected	Observed	$p <$	Expected	Observed	$p <$
Gibbs & Ng (1979)						
Figure 1	16	4.7		22	5.3	
Figure 2	14	3.2		20	3.1	
Figure 3	12	3.2		19	9.2	
All figures	42	11.1	.00001	61	17.6	.00001
Gibbs & Ng (1984a)						
Figure 1, top	32	12.3		46	40.4	
Figure 1, bottom	14	1.0		21	6.0	
Both figures	46	13.3	.00001	67	46.5	.03
Gibbs & Ng (1984b)						
Figure 1	29	3.7		34	26.0	
Figure 2	20	4.9		31	22.4	
Figure 3	26	10.4		37	20.4	
Figure 4	16	11.0		26	15.2	
Figure 5	28	16.4		41	19.9	
Figure 6	16	4.8		26	22.9	
All figures	135	51.2	.000001	195	126.7	.0001

Note. The p values are the likelihood that the chi-square values would be as low or lower by chance.

chi-squares, and all six indicate that variability was reliably less than expected by chance. In five of the six cases, $p < .0001$. For a similar use of chi-square tests, see Fisher (1936).

The data are also remarkable in other ways. The narrowness of the fissures, especially the later fissure, shows a precision of timing (when precision is measured by relative error) that far exceeds anything I have seen in other behavioral data involving the timing of durations. For example, the second fissure in Figure 1 is centered at 80 min, and its width halfway between top and bottom is about 10 min. The ratio of width to center is about 0.13. In contrast, experiments by others that have found dips in retention have observed ratios of width to center that are always roughly 1.0 or more (Cherkin, 1971; Frieder & Allweis, 1978, 1982; Irwin, Banuazizi, Kalsner, & Curtis, 1968; Menzel, 1979; Zerbolio, 1969; and many studies of the Kamin effect, e.g., Kamin, 1957). In other experiments involving duration timing, ratios of width to center are roughly 1.0 or more (e.g., Church & Gibbon, 1982; Gormezano, Kehoe, & Marshall, 1983; Heinemann, 1974; Roberts, 1981). Duration timing (done by egg timers and stopwatches) should not be confused with cycle timing (done by alarm clocks and wall clocks); cycle timing in animals is sometimes extremely precise (e.g., DeCoursey, 1960). A second remarkable feature of the data is that there are *two* dips in retention. Many labs, using a variety of animals and tasks, have observed a single dip in the function relating retention to TTI (e.g., the references given above), but no other lab, as far as I know, has found two.

Drug Effects

The first evidence for the three-stage model came from the study of drugs that produce amnesia. The article that proposed the three-stage model (Gibbs & Ng, 1976) contains two figures, one of which is reprinted here as Figure 2. It shows the effect of different drugs (saline, ouabain, LiCl, KCl) on the learning task discussed earlier. The two lines in the ouabain panel represent results from different injection times; the two lines in the saline and LiCl represent replications (K. T. Ng, personal communication, March 18, 1985). The text states that there was "a different group of 20 chickens at each learning-retention interval." I will assume that different points are based on different groups of 20 chicks; in other papers (e.g., Gibbs & Ng, 1984a), this is made explicit.

The data of Figure 2 show the anomaly noted earlier, less-than-expected variability. When there is more than one point at a given time of testing, the two points are on average too close together. As before, this can be measured with chi-square tests. Each pair of points generates a 2×2 table, where the two rows are the two possible responses (avoid or not avoid) and the two columns are the two points. The chi-squares, with 1 degree of freedom per table, can be added up across tables. Unlike the data of Figure 1, the data of Figure 2 were not always easy to convert to numbers because the points were not always centered on a percentage divisible by 5. When two points overlapped, I assumed that they represented the same value; otherwise, in cases of doubt, I chose the percentage that gave the most variability. Chi-square values for the three panels of Figure 2 with repetitions, each

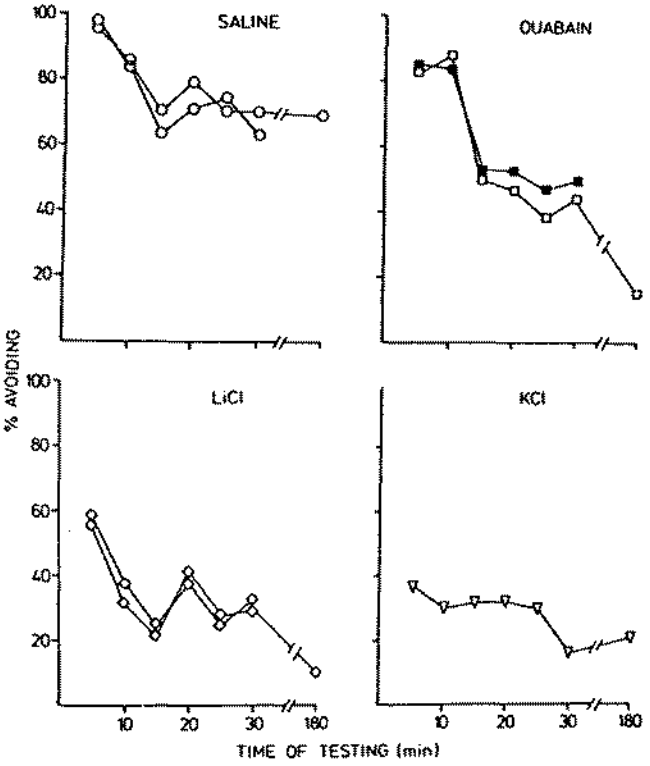

Figure 2. Retention as a function of time between training trial and test trial. (Different panels correspond to different drugs given before the training trial.) *Note.* From "Memory Formation: A New Three-Phase Model" by M. E. Gibbs and K. T. Ng, 1976, *Neuroscience Letters, 2*, Figure 1, p. 167. Copyright 1979 by Elsevier/North-Holland. Reprinted by permission.

with an expected value of 6, are 0.9 (saline, $p < .05$), 1.1 (ouabain, $p < .05$), and 0.7 (LiCl, $p < .01$). The overall chi-square is 2.7 with an expected value of 18; the probability that a value as low would occur by chance is less than .001 The other figure of Gibbs and Ng (1976) contains no repeated points, so the same analysis cannot be done.

Less-than-expected variability occurs yet again in a long paper (Gibbs & Ng, 1977) that reviews old evidence and presents new evidence. The paper contains 15 figures with data; I consider here only some of the data from two of them (Figures 5 and 8 of Gibbs & Ng, 1977, reprinted here as Figures 3 and 4). The two chosen figures share a procedural feature not shared by any of the other 13 figures, but the selection should be taken into account when judging the p values. The experimental procedure is the same as before, with each point based on a different group of 20 chicks. The different lines, connecting different plotting symbols, correspond to different times that a amnesic agent (Figure 3, ouabain; Figure 4, cycloheximide) was injected, ranging from 30 or 45 min before the training trial to 10 or 30 min after the training trial.

Leaving aside (a) in Figure 3 only, the two extreme times of injection (30 min before and 10 min after), which seem systematically different from the rest, and (b) the results from TTIs of more than 30 min, where the lines seem to diverge

systematically, the rest of the data seem too close together. In each case, the lines follow too closely a common path. The likelihood that the lines would be so close by chance can be measured just as it was measured with Figure 2. Each TTI generates a table and a chi-square value; for example, if there are 4 points at one TTI, this generates a 2 × 4 table. Adding up chi-squares across TTIs generates a chi-square for each figure. The data of Figure 3 generate a chi-square of 6.3 with an expected value of 17, $p < .01$. The data of Figure 4 give a chi-square of 3.5 with an expected value of 18, $p < .01$. The overall chi-square of the two figures taken together is 9.8 with an expected value of 35, $p < .0001$. This p value may be artificially low because of the selection of data; a more conservative assessment would be the maximum of the p values for each figure alone (< .01).

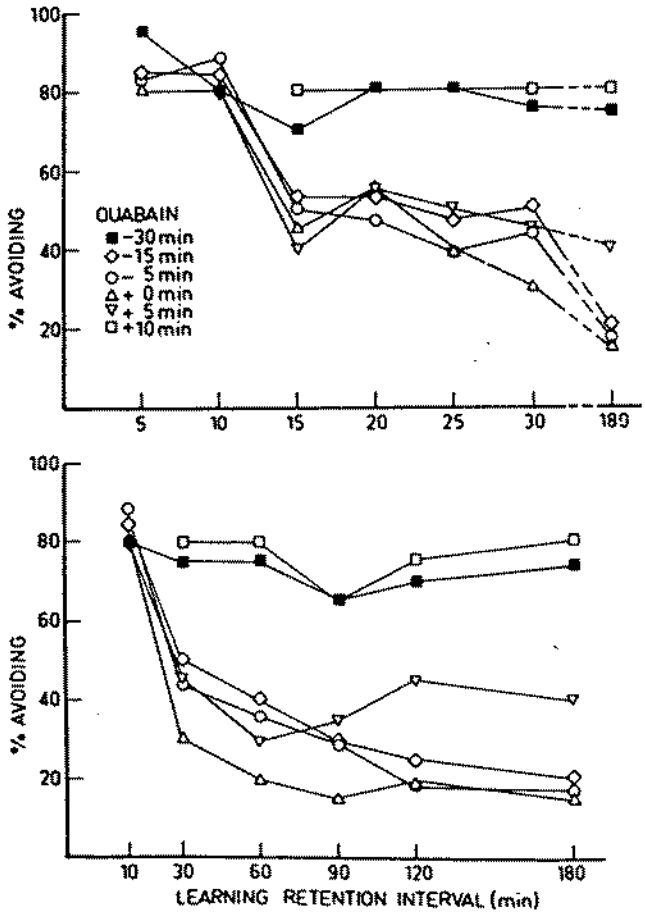

Figure 3. Retention as a function of time between training trial and test trial. (Different lines correspond to different times of injection of ouabain, from 30 min before the training trial to 10 min after.) *Note.* From "Psychobiology of Memory: Towards a Model of Memory Formation" by M. E. Gibbs and K. T. Ng, 1977, *Biobehavioral Reviews, 1*, Figure 5, p. 122. Copyright 1977 by ANKHO International. Reprinted by permission.

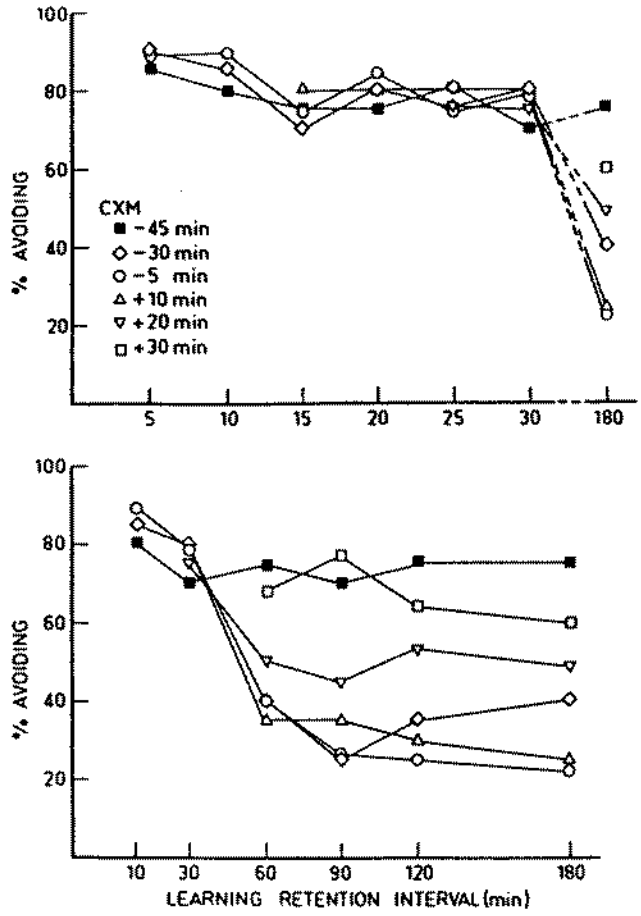

Figure 4. Retention as a function of time between training trial and test trial. (Different lines correspond to different times of injection of cycloheximide, from 45 min before the training trial to 30 min after.) *Note.* From "Psychobiology of Memory: Towards a Model of Memory Formation" by M. E. Gibbs and K. T. Ng, 1977, *Biobehavioral Reviews, 1*, Figure 5, p. 124. Copyright 1977 by ANKHO International. Reprinted by permission.

Discussion

In five papers, the variability of some or most of the data is less than expected if (a) there were 20 chicks per datum and (b) trials with different chicks were independent. In most cases (e.g., Figure 1), the spacing of the data points leaves no doubt about the sample size. As for the assumption of independence, there are four ways that it could be violated: either within groups or between groups, and with either a positive or a negative correlation. However, only one of the four possibilities would produce the less-than-expected variability detected here: a within-groups negative correlation, which means that if one chick avoids, then other chicks in the group are less likely to avoid. The most obvious source of lack of independence, fluctuations in the experimental conditions, would produce a *positive* correlation between chicks. An assumption of independence is made every time a test of significance is used,

and there is nothing in the published descriptions of Gibbs and Ng's procedure to make the assumption of independence less plausible than usual.

In a letter, however, Ng states that chicks were tested in pairs, a fact not made clear in any of the papers (personal communication, March 18, 1985). This raises more strongly than usual the possibility of dependence between chicks. Chi-square tests can be used to ask if a negative correlation within pairs of chicks tested together could produce the less-than-expected variability. Consider Table 2. It gives the probability of avoidance for the second-tested chick of a pair contingent upon what the first-tested chick did. The correlation between chicks is as strong as possible—in the negative direction—given that both chicks have an overall probability of avoidance of .8. Table 2 should help make it clear that when the correlation between chicks within pairs is as negative as possible, each *pair* of chicks produces only one of two results. When the overall probability of avoidance is greater than .5, the two results are (a) one chick avoids or (b) both chicks avoid. Thus, with 20 chicks (10 pairs), the total number of avoidances is between 10 and 20. When the overall probability of avoidance is less than .5, the two results are (a) neither chick avoids or (b) one chick avoids.

Thus, with maximum negative correlation, a total of 17 avoidances out of 20 trials (10 pairs of chicks) means that in 7 pairs, both chicks avoided; in 3 pairs, one chick avoided. We can now do chi-square tests just as before except that the tested tables contain counts of *pairs* of chicks rather than counts of individual chicks (e.g., 17 and 3 are replaced by 7 and 3). These tests assume the maximum possible negative dependence within pairs and independence between pairs. The chi-squares based on these new assumptions are still less than their expected values in all 27 cases considered earlier (22 in Table 1, 3 in Figure 2, 1 in Figure 3, and 1 in Figure 4), and the p values are still very low. For the data of Table 1, for example, five of the six overall p values are less than .002. So dependence within pairs cannot explain the lack of variability. This is consistent with Ng's statements that "examination of the raw data does not substantiate lack of independence between trials" and "whatever evidence of lack of independence there is, it would seem to be insufficient to account for the phenomenon under discussion" (personal communication, March 18, 1985).

Table 2
Example of Maximum Negative Dependence Within a Pair of Chicks

	Probability that second chick	
Action of first chick	Will avoid	Will not avoid
Avoided	.75	.25
Did not avoid	1.0	0

Note. Entries in the table are probabilities of avoidance for the second-tested chick in a pair. The correlation between chicks is as negative as possible assuming that the overall probability of avoidance is .80 for both chicks.

The evidence discussed here is not the only evidence for the Gibbs and Ng model. Gibbs and Ng have reported other evidence, involving drug effects, not considered here (e.g., Gibbs, 1983; Gibbs & Ng, 1978), and others have described evidence for three-or-more-stage models of memory formation (e.g., Frieder & Allweis, 1978, 1982) that are somewhat different than the Gibbs and Ng proposal. But the articles discussed here contain some of the most important evidence. Until it is explained, the less-than-expected variability in these papers should reduce one's belief in the model.

References

Cherkin, A. (1971). Biphasic time course of performance after one-trial avoidance learning in the chick. *Communications in Behavioral Biology, 5*, 379–381.

Cherkin, A., & Lee-Teng, E. (1965). Interruption by halothane of memory consolidation in chicks. *Proceedings of the Federation of American Societies for Experimental Biology, 24*, 328.

Church, R. M., & Gibbon, J. (1982). Temporal generalization. *Journal of Experimental Psychology, 8*, 165–186.

DeCoursey, P. J. (1960). Daily light sensitivity rhythm in a rodent. *Science, 131*, 33–35.

Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Annals of Science, 1*, 115–137.

Frieder, B., & Allweis, C. (1978). Transient hypoxic-amnesia: Evidence for a triphasic memory-consolidating mechanism with parallel processing. *Behavioral Biology, 22*, 178–189.

Frieder, B., & Allweis, C. (1982). Memory consolidation: Further evidence for the four-phase model from the time-course of diethyldithiocarbamate and theacrynic acid amnesias. *Physiology and Behavior, 29*, 1071–1075.

Gibbs, M. E. (1983). Memory and behaviour: Birds and their memories. *Bird Behaviour, 4*, 93–107.

Gibbs, M. E., & Ng, K. T. (1976). Memory formation: A new three-phase model. *Neuroscience Letters, 2*, 165–169.

Gibbs, M. E., & Ng, K. T. (1977). Psychobiology of memory: Towards a model of memory formation. *Biobehavioral Reviews, 1*, 113–136.

Gibbs, M. E., & Ng, K. T. (1978). Memory formation for an appetitive visual discrimination task in young chicks. *Pharmacology Biochemistry & Behavior, 8*, 271–276.

Gibbs, M. E., & Ng, K. T. (1979). Behavioral stages in memory formation. *Neuroscience Letters, 13*, 279–283.

Gibbs, M. E., & Ng, K. T. (1984a). Diphenylhydantoin extension of short-term and intermediate stages of memory. *Behavioural Brain Research, 11*, 103–108.

Gibbs, M. E., & Ng, K. T. (1984b). Hormonal influence on the duration of short-term and intermediate stages of memory. *Behavioural Brain Research, 11*, 109–116.

Gormezano, I., Kehoe, E. J., & Marshall, B. S. (1983). Twenty years of classical conditioning research with the rabbit. *Progress in Psychobiology and Physiological Psychology, 10*, 197–275.

Heinemann, E. G. (1974). Multimodal distributions of pigeon's reaction time. *Bulletin of the Psychonomic Society, 3*, 75–77.

Irwin, S., Banuazizi, A., Kalsner, S., & Curtis, A. (1968). One-trial learning in the mouse: I. Its characteristics and modification by experimental-seasonal variables. *Psychopharmacologia, 12*, 286–302.

Kamin, L. J. (1957). The retention of an incompletely learned avoidance response. *Journal of Comparative and Physiological Psychology, 50*, 457–460.

Lee-Teng, E., & Sherman, S. M. (1966). Memory consolidation of one-trial learning in chicks. *Proceedings of the National Academy of Sciences (U. S. A.), 56*, 926–91.

Menzel, R. (1979). Behavioural access to short-term memory in bees. *Nature, 281*, 386–389.

Roberts, S. (1981). Isolation of an internal clock. *Journal of Experimental Psychology: Animal Behavior Processes, 7*, 242–268.

Zerbolio, D. J. (1969). Memory storage: The first posttrial hour. *Psychonomic Science, 15*, 57–58.

Received December 29, 1984
Revision received January 13, 1986 ■