THE RELIABILITY OF RAT LEARNING SCORES FROM THE MULTIPLE-T MAZE AS DETERMINED BY FOUR DIFFERENT METHODS[1]

CALVIN P. STONE, *Stanford University,*
AND
DOROTHY BIRD NYSWANDER, *University of Utah*

The purpose of this report is three-fold: (1) to present and compare four methods of calculating reliability coefficients for maze scores; (2) to determine the most reliable segments of the total series of trials to which the animals have been subjected; and (3) to indicate the effect of increasing the number of subjects on the stability of the reliability coefficients for different parts of the trial series.

Since Maupin (1921) and Tolman and Nyswander (1927) have summarized the literature on the choice and evaluation of learning criteria and the reliability of the measures adopted, we shall not deal with this phase of the subject as a whole, but, in the body of the paper, shall touch upon such topics as are pertinent to our present interests.

TECHNIQUE

The size of reliability coefficients for maze data depends not only upon the methods by which they are computed but also upon the type of maze used, the sample of the animal population, motivation of the learners, learning scores selected, etc.; hence we shall discuss the latter topics in considerable detail at the outset.

1. *Apparatus*: The Multiple-T Maze

The most important characteristic of this maze may be described by saying that every true-path leads to a cross-road; here the next true-path leads off at right angles in one direction and a blind alley leads off at right angles in the other direction. This principle of construction may be noted in the floor plan, Fig. 1. A., and the detail, Fig. 1. B., illustrating the region of the starting box.

The walls are 4 inches high and the alleys 4 inches wide. Ideally, blind alleys and true pathways are intended to be

[1]This is the first of a forthcoming series of articles dealing with the age factor in animal learning. The research was financed by a grant from the Carnegie Corporation.

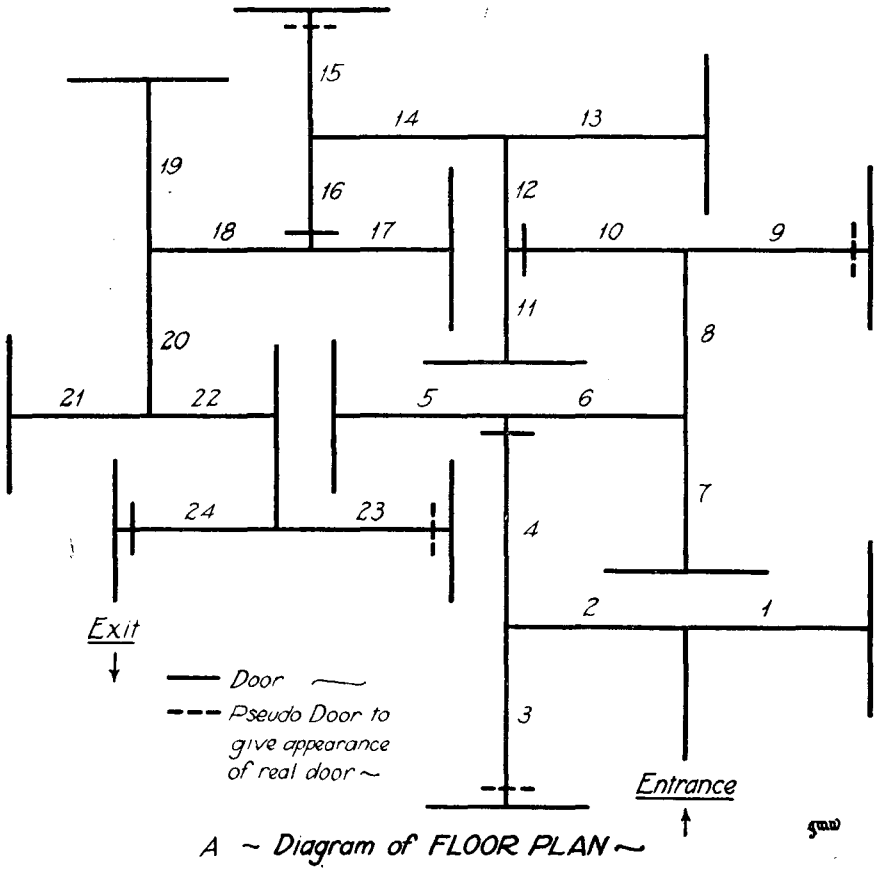

FIGURE 1. A.
FLOOR PLAN OF THE MULTIPLE-T MAZE

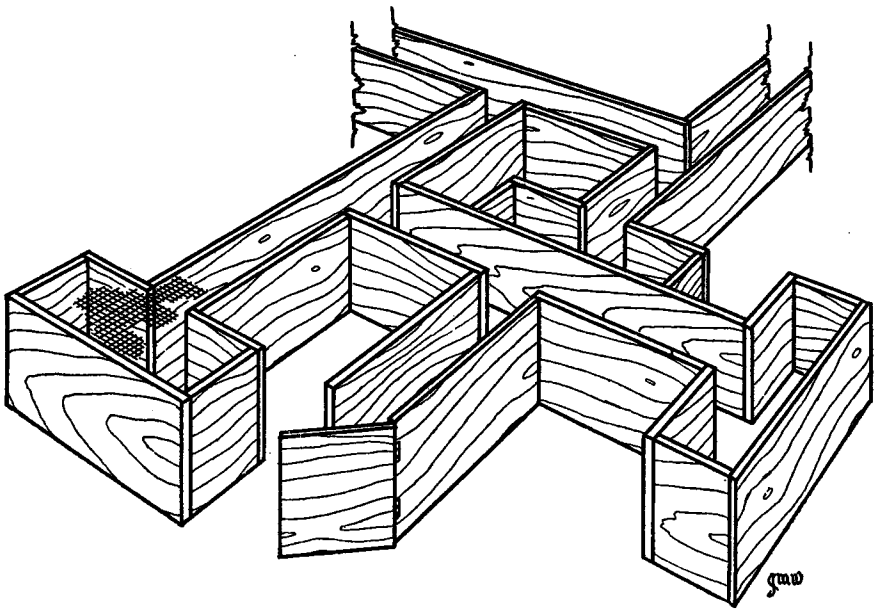

FIGURE 1. B.

of equal length but practical considerations of construction caused slight variations to be made. These, however, are far below the amount necessary for discrimination of pathways on the basis of distance traversed (Yoshioka 1926). The length of the usual pathway, whether blind alley or true-path, is 16 inches, with some exceptions as indicated in Fig. 1. A. To make blind and true paths appear alike to the animal making a choice of direction, a pseudo cross-road is put at the distal end of each blind. This is sufficiently long to keep the animal from seeing that it is really only a blind road until he goes out to the end of the alley. To the experimenters it seemed probable that this device would enhance the difficulty of each choice of direction, since the animal, standing on the threshold of the cross-road, could not see that one road had a blind ending and the other led to a new pathway. So far no special experiments have been made to test our theory, and nothing in our routine experiments or observations clearly confirms or refutes the assumption. Settlement of this point, however, is not necessary for our present interest, since all of our animals were subjected to the same conditions.

A door prevents the animal's return into the starting box after it has entered upon the first pathway of the maze proper. At four points in the maze, doors are so placed at the distal ends of true pathways that retracings beyond these doors are impossible (See Fig. 1. A.). To make all pathways containing doors alike at each end, pseudo-doors were put at the distal ends of the alleys in positions opposite the real doors. These, of course, are immobile and serve only to eliminate the possibility of a choice on the basis of the presence or absence of a door. The doors are constructed of hardware cloth attached to a light wire frame and are operated by strong threads centralized near the experimenter's chair. They hinge from the top of the maze, and when released drop down into the pathway. When open they fit up snugly against the wire mesh covering the maze.

The floor is not attached to the maze box. It consists of battleship linoleum to which a coat of varnish has been applied. After each day's work the floor is washed with clear water.

A hanging lamp suspended from the ceiling of the room at a height of approximately 10 feet furnishes the illumination for the maze. The light source is a 175 Watt, 110 Volt, Mazda lamp. Owing to the height of the lamp and the diffuse

illumination provided by the milk glass reflector and globe about the lamp, shadows in the maze are totally absent in almost all of the pathways and are very dim in the remainder. If present they are exactly the same in the true path and alley of any unit.

2. Animals

The animals used in this study were reared in our own laboratory and, without exception, were of known ages. Their diet before and during the learning period consisted of the dry food mixture given below. Since this is a well balanced diet one may consider the dietary deficiency during the learning period, when hunger is the chief motive for maze running, as falling chiefly within the class of *quantitative* rather than *qualitative* deficiency.

STANDARD DIET (MCCOLLUM)

Whole wheat, ground fine	67.5
Casein	15.0
Whole milk, powdered	10.0
Calcium Carbonate	1.5
Sodium Carbonate	1.0
Butter fat	5.0
Total	100.0

Prior to being put on the learning experiment the animals were allowed to eat ad libitum; hence they were large in size even at the earliest ages considered in the experiment. In fact, for the most part, they had passed the age at which the most rapid gain in body weight takes place. Table I gives the animal groups and other pertinent data concerning each group: namely, age at which learning experiments began, number of litters represented, and number of individuals of each sex.

3. General Technique for Conduct of the Experiments

a. Preliminary Experiments: Before the animals were started on the maze they were given five trials, one trial per day, on a simple platform escape box. The details of this apparatus need not concern us here. Suffice it to say that the hungry animal is put into the box and if he steps on a small platform in the center of the floor an electric circuit is closed thereby activating a magnet which withdraws the bolt holding the door. When the bolt is withdrawn, the door automatically opens and permits the animal to enter the food box just outside the door.

Preliminary training is undoubtedly a desirable feature of all animal experiments, as has been demonstrated by the various experiments on transfer of training in rats and the specific experiment by Warden (1925) covering this point. In addition to familiarizing the animal with the novel situation involved in the experiment itself, putting him in a problem solving

TABLE I

ANIMAL GROUPS WITH THEIR RESPECTIVE AGES, NUMBERS OF LITTERS, AND NUMBERS OF INDIVIDUALS DISTRIBUTED AS TO SEX. OBVIOUSLY, CONTROL GROUPS HAVE THE SAME NUMBERS OF LITTERS AS THE GROUPS OF WHICH THEY ARE THE CONTROLS.

Age group	Number of litters	Total number of individuals		
		Male	Female	Total
75 days	7	12	13	25
5 months	8	7	9	16
6 months	9	18	11	29
8 months (ctr. 75 day gr.)		11	14	25
9 months	7	13	12	25
12 (a) mo. (ctr. 6 mo. gr.)		11	14	25
12 (b) months	9	16	19	35
18 mo. (ctr. 12 (b) mo. gr.)		12	13	25
Total no. different litters	40	100	105	205

state, etc., it gives the experimenter an opportunity to study individual animals for the purpose of determining which, if any, need a great amount of handling to make them thoroughly gentle and tame, which should have their ration drastically limited in order to keep them from putting on fat during the experiment, and which animals, if any, will require an extra allotment of food in order to keep them in good physical condition throughout the experimental series. (In our experiments some animals were under training three months or more before going back on an unrestricted diet.) It also gives the animals opportunity to adjust to a feeding schedule of one meal in 24 hours as opposed to five or six meals as is its custom. (Wang, 1925)

b. Control of Diet: For a period of 24 to 36 hours immediately preceding the first trial of preliminary training the animals were deprived of food for the purpose of insuring the proper degree of hunger and activity for the first day's trials. During the next 10 to 15 days they received as a general rule only the amount of dry food consumed in a feeding

period of twenty minutes. If individuals were relatively heavy for their respective ages or relatively thin (rare), deviations from this general rule were made to suit the cases. At the end of 10 to 15 days (this being the fifth to the tenth day on the maze) a supplementary allowance of from 1 to 3 grams per individual was given after the animals had been returned to their cages. Our purpose was to hold the adult animals at maintenance after the initial period of 10 to 12 days and to allow the immature animals to grow only slightly. To hold young rats, in which the growth tendency is strong, at maintenance, one must deprive them of food much more drastically than is the case with mature animals. They can be maintained, however, on a diet that permits of a slight daily increase in weight (1½ to 2 grams for very young learners, and ½ to 1 gram for animals from 2 to 3 months of age) without appreciable loss of motivation from hunger.

Each animal was weighed daily before its learning experiment and the weight recorded along with its other records. At the time of the weighing, the experimenter made all necessary decisions as to the advisability of a special variation from the routine method of rationing the animal.

c. Use of Doors in Maze: The positions of the doors in the maze have already been indicated. Their use can be briefly stated. In the early trials all doors were quietly closed in the wake of the animal to prevent its retracing into the pathway just travelled. As will be seen from observing the positions of the doors in Fig. I-A, we have permitted the animal a little lee-way for retracing for exploratory purposes without permitting him to make long backward runs as is permitted by the older techniques dating back to the methods of Small (1899) and Watson (1903). The latter methods have very few points to recommend them[2] and many to justify their being discarded. As soon as the forward orientation habit is set up it becomes unnecessary to close doors after the animal as it will not ordinarily return to them. In case it reverses its direction, however, the door is quickly closed to prevent retracing beyond that point.

d. Time and errors: The time for a given trial consists of the interval between the animal's *leaving* the starting box and its *arriving* at the door of the exit. No static time for any purpose whatever was removed from this time interval.

[2]For critical discussion of this point see the article by Tolman and Nyswander, 1927.

Two types of errors were recorded. Type I consisted of entrances into blind alleys as the animal was oriented toward the goal. An error was recorded if it progressed into the blind alley with approximately two-thirds or more of its body length. Slight turnings in the direction of the blind, mere gestures as it were, were not recorded as errors although they sometimes appeared to be erroneous moves inhibited after the initial step had been taken. They are of relatively rare occurence for a given run although in a half day's work the experimenter may observe them many times.

Type II errors consist of short backward runs on the true pathway or into blind alleys. These drop out almost entirely within the first five to ten trials in our experiments and are of relatively infrequent occurrence even in the early trials, owing to the fact that the closed doors facilitate setting up the habit of forward orientation.

e. Choice of Error Scores for Study of Reliability: The choice and validation of scores in maze studies by which learning ability is measured has been reviewed by Maupin and critically examined by Tolman and Nyswander (1927). Of the various criteria, time, entrances into blind alleys, retracings, and number of perfect runs in a given series, the latter reviewers consider that elimination of entrances into blind alleys is the best single measure of learning ability.[3] With the present status of maze technique this criterion of learning appears to us to be somewhat more satisfactory, all things considered, than any of the other aforementioned criteria. Hence we have used only errors of Type I in this study. Errors of Type II occur so infrequently after the first few trials that they are useless for a study of this type.

4. Four Methods of Determining the Reliability of Error Scores

The methods by which error scores were correlated to determine their reliability for different series of trials may be briefly described as follows:

METHOD A. Correlation of the sum of the errors on the odd trials with the sum of the errors on the even trials.

This method gives an index of the smoothness of the individual learning curves or the degree to which fluctuations from

[3]No attempt has been made, so far as we are aware, to evaluate the errors made in the individual blinds of a maze in terms of their relative efficacy in measuring animal learning. Such a study is now being made by the authors of this paper.

the general trend of the performances appearing on odd days tend to balance those on even days.

METHOD B. Correlation of the sum of errors on the odd numbered blinds with the sum of errors on the even numbered blinds, i. e., sum of errors on blinds 1, 5, 9, 13, 17, 21, with the sum of errors on blinds 3, 7, 11, 15, 19, 23. (See Fig. 1. A.)

This and the method which immediately follows is analogous to the procedure followed in the construction of certain mental and educational tests. The blinds correspond to the different test items. In computing the reliability they are divided into halves which, it is assumed, give independent samplings of the ability of the animals in the function measured.

METHOD C. Correlation of the sums of all errors for the first half of the maze with the sums of all errors for the second half of the maze, i.e., sums of errors on blinds 1-11 with sums of errors on blinds 13-23. (See Fig. 1. A.)

Roughly stated, it is assumed that the animal has run two adjacent mazes without interruption of progress from one to the other. This is analogous to the use of two forms of the same test when dealing with the problem of reliability in test construction. In this case the two mazes are identical as to mode of construction and number of blinds, but different in the specific requirements for individual choices.

METHOD D. Correlation of the sums of errors for any segment of the trial series with the sums of errors for any other segment of the trial series, i.e., sum of the errors for trials 1-10 with those of trials 11-20, sum of errors for trials 6-10 with those of trials 11-15, etc.

High correlations are obtained by this method only when the animals' learning performances are quite consistent for the two segments correlated, a condition that exists only when the individual learning curves exhibit a certain parallelism in direction within the segments. For studies in which knowledge of relative placements of animals is desired, such as a study of individual differences or family similarities, this method of correlating scores is especially important because it indicates the degree to which performances of individuals remain consistent for various segments of the trial series.

When determining reliability by Method D, note must be taken of the trials yielding zero errors (perfect runs). Such scores occurring at the end of the learning series serve in the case of segmental correlations, such as trials 1-15 vs.

16-30, to introduce scores in the second half that lessen the variability and thus lower the correlation coefficient. The zero scores are in themselves indeterminate values of the ability measured. For the first three methods of determining reliability the coefficients are not appreciably affected by the introduction of a relatively small number of zero scores, since the effect is that of adding a constant to both variables correlated.

5. Methods of Determining the Reliability of Time Scores

In the present study we have centered our attention upon the reliability of error scores, but for purposes of comparison coefficients for time scores have been computed for certain groups of animals. Since time was not recorded for each maze element, as Methods B and C demand, the nature of our available time scores makes it possible to use only Methods A and D to determine their reliability. Without much difficulty, however, were it desirable, one might collect records by which Method C could be used. To do this one would have to use two watches, or one with two second hands, so that the time required by the animal to run past the half-way point in the maze as well as to complete the run through the entire maze might be recorded. Up to the present time we have not had occasion to collect such records systematically, but a few trials have convinced us of the feasibility of the method.

Coefficients of Reliability for Errors

The coefficients of reliability for the individual groups of animals are presented in Tables 2-5. The coefficients of Table 2 were obtained by correlating the sums of errors for odd trials with the sums of errors for even trials (Method A); those of Table 3, by correlating the sums of the errors made on odd blinds with the sums of errors made on even blinds (Method B); those of Table 4, by correlating the sums of errors made on the first half of the maze with the sums of errors made on the second half of the maze (Method C); and those of Table 5, by correlating the sums of all errors for one segment of the series of trials with the sums of all errors for another segment of trials (Method D.). The organization of the data in each of the tables is probably self-explanatory or can be readily understood with a brief explanation. Note Table 2. The coefficients of reliability for various segments of the trial series between 1 and 30 are listed from left

to right for each of the groups of animals studied. Reading any of the columns from top to bottom one may compare the magnitude of coefficients for the different groups of animals as obtained from a given segment of trials, e.g., 1-10, 11-20, etc.

The coefficients appearing in these tables were corrected by the Spearman-Brown formula[4] for halving of the data. Their probable errors were computed from the formula derived by Shen (1926)[5] for coefficients of reliability obtained from the use of the Spearman-Brown formula. The applicability of this correction seems justified in view of its analogous usage in the field of psychological tests and measurements. Neither the fact that the homogeneity of our groups of animals is an unknown factor nor the fact that we are correlating the results of a test running through a considerable period of time is overlooked in this connection. They do not alter the reasonable assumption that a better estimate of the reliability of our data is obtained through the use of this formula than by failure to do so.

Tables 2, 3, and 4 are presented in sequence so that the coefficients obtained by the three methods of calculating reliability may be compared and, in the next section of this paper, may be contrasted for different segments of the trial series. A survey of the coefficients shows that they are relatively high. This is especially noteworthy because coefficients obtained by previous investigators have been comparatively low as a rule, and in no case uniformly high. (See paper by Tolman and Nyswander, 1927.) With but few exceptions the coefficients of Table 2, obtained by correlating the sums of errors for odd trials with the sums of errors for even trials (Method A), are the highest. Those of Tables 3 and 4 are quite similar for the respective columns, although in a majority of instances the coefficients of the former are slightly higher than those of the latter. The similarity of the coefficients obtained by Methods B and C confirms our expectations if we assume that the blinds are analogous to test elements in a reliable test. Before making these calculations we were unable to say whether the elements to be learned were so distributed throughout the maze as to give reliable coefficients. It seemed entirely possible that the effect of certain positions of blinds or sequence

[4]Spearman-Brown formula: $R = \frac{nr}{1+(n-1)r}$

[5]Shen's formula for $P.E._R$: $P.E._R = \frac{2(1-r^2)}{\sqrt{N}(1+r)^2}$

TABLE 2

COEFFICIENTS OF RELIABILITY COMPUTED BY METHOD A (ODD TRIALS VS. EVEN TRIALS)

Group	N	Segments of Trials					
		1-30	1-20	11-30	1-10	11-20	21-30
75 days	25	.913±.011	.851±.020	.919±.011	.824±.027	.883±.017	.947±.008
5 months	16	.985±.002	.974±.004	.980±.003	.857±.025	.974±.004	.969±.005
6 months	29	.974±.002	.857±.018	.980±.002	.802±.027	.958±.005	.930±.010
8 months	25	.964±.005	.936±.008	.969±.005	.780±.033	.942±.008	.895±.012
9 months	25	.953±.005	.901±.012	.969±.005	.734±.040	.919±.011	.969±.005
12 (a) months	25	.947±.008	.930±.001	.964±.005	.851±.020	.919±.011	.817±.027
12 (b) months	35	.942±.007	.864±.017	.964±.005	.630±.053	.925±.009	.953±.005
18 months	25	.913±.011	.857±.020	.990±.002	.592±.067	.947±.008	.947±.008

TABLE 3

COEFFICIENTS OF RELIABILITY COMPUTED BY METHOD B (ODD BLINDS VS. EVEN BLINDS)

Group	N	Segments of Trials					
		1-30	1-20	11-30	1-10	11-20	21-30
5 months	16	.947±.009	.919±.014	.936±.011	.930±.012	.870±.023	.958±.007
6 months	29	.953±.005	.919±.010	.936±.007	.780±.030	.870±.019	.913±.010
8 months	25	.870±.020	.901±.012	.780±.049	.857±.020	.788±.033	.710±.047
9 months	25	.930±.011	.953±.005	.837±.023	.901±.012	.844±.023	.810±.030
12 (a) months	25	.851±.020	.817±.027	.780±.049	.765±.036	.795±.030	.773±.049

TABLE 4

COEFFICIENTS OF RELIABILITY COMPUTED BY METHOD C (FIRST ½ OF MAZE VS. SECOND ½ OF MAZE)

Group	N	Segments of Trials					
		1-30	1-20	11-30	1-10	11-20	21-30
5 months	16	.901±.017	.953±.008	.942±.010	.851±.027	.936±.011	.913±.015
6 months	29	.964±.005	.901±.013	.930±.010	.851±.018	.851±.018	.901±.013
8 months	25	.925±.011	.844±.023	.844±.023	.773±.033	.788±.033	.693±.048
9 months	25	.817±.027	.773±.033	.817±.027	.658±.055	.765±.036	.773±.033
12 (a) months	25	.857±.020	.844±.023	.830±.027	.964±.005	.773±.049	.837±.023
18 months	25	.953±.005	.883±.017	.964±.005	.780±.033	.913±.011	.925±.011

TABLE 5

COEFFICIENTS OF RELIABILITY COMPUTED BY METHOD D (SEGMENT VS. SEGMENT)

Group	N	Segments of Trials					
		(1-15) vs. (16-30)	(1-10) vs. (11-20)	(1-5) vs. (16-20)	(6-15) vs. (16-25)	(11-15) vs. (16-20)	(21-25) vs. (26-30)
5 months	16	.947±.009	.901±.017	.750±.048	.958±.007	.942±.010	.969±.005
6 months	29	.851±.018	.889±.016	.773±.030	.883±.016	.930±.010	.824±.024
8 months	25	.765±.036	.810±.030	.462±.099	.947±.008	.876±.017	.870±.020
9 months	25	.936±.008	.919±.011	.817±.027	.883±.017	.942±.008	.592±.067
12 (a) months	25	.693±.048	.693±.048	.571±.072	.795±.030	.857±.020	.726±.044
18 months	25	.750±.040	.857±.020	.485±.095	.857±.020	.876±.017	.844±.023

of turns in the maze, i.e., in the middle, near the entrance or food box, etc., would serve to lower the correlations obtained by either or both of the methods. Hence it was quite gratifying to find that there is a sufficient number of elements in the Multiple-T maze to differentiate the behavior of the animals consistently and that the resulting coefficients of reliability are sufficiently high for our present experimental purposes.

Table 5 gives the coefficients obtained by correlating the scores made in one part of the learning series with scores made on various other parts of the series. If columns 1 and 2 of this table are compared with columns 1 and 2 of Tables 2, 3, and 4, it is seen that with one exception (12 (a) months group) the coefficients for corresponding groups of the former are somewhat lower. It is more difficult to obtain high reliability coefficients by this method than by any of the others. This is readily understood, however, when it is remembered that a high reliability coefficient obtained by this method means that there is little or no crossing over of the individual learning curves.

At the present time we cannot establish with a high degree of finality the relative merits of the foregoing methods of computing reliability. Ultimately, however, this may be done by determining their relations to an outside criterion such as the intercorrelations between scores for a given group of animals running successively on different mazes. If we conceive of "ability to learn a maze" as a function that exists in varying amounts from animal to animal and depends for its accurate measurement on the reliability of the maze and maze technique, we should secure a fairly accurate measure of this ability by running them on several different mazes. Through indirect relationships between these measures of ability and the reliability coefficients of the mazes as determined by each of the four methods we may then be able to say which of them most accurately measures the reliability of the learning scores.

As we have stated heretofore (pp. 504-6) the four methods give us rather specific information concerning the consistency of the animal's performances from day to day, the nature of the maze as a differentiator of behavior at different points, and the consistency with which errors appear in the first and the second halves of the maze. Because of their seeming specificity of significance we should recommend for the present that reliability of learning scores be determined for any set of data by two or more of the four methods. When information is

desired concerning the degree to which animals hold their relative rankings for error elimination through different segments of the learning series one of the methods employed should be Method D. Eventually further information may make possible the prediction of reliability for one method in terms of the reliability calculated by another.

The Reliability of Different Segments of the Trial Series

Hunter (1922, 1924), Tolman and Nyswander (1927), and Burlingame (1927) have attempted to find the most reliable part of the trial series for their respective mazes. Hunter was interested in the problem *per se*; the others sought this information in order that the scores used in their investigations might be taken from the most reliable portion of the trial series. Tolman's highest reliabilities were obtained by discarding the first two or three trials and using the next 10 to 30 trials. Burlingame found trials 6-15 somewhat the best of a small number of trials taken in the series where error elimination was proceeding at the most rapid rate on the Multiple-T maze. Other portions of the series based on trials beyond the first five also gave reliability coefficients ranging from 0.75 to 0.95.

In this portion of our paper we are concerned with the relative merits of various segments of the total trial series from the standpoint of the size of their reliability coefficients as calculated by the four methods. Comparisons are based on the coefficients reported in Tables 2-5.

The initial step in making this comparison consisted of averaging the coefficients in each of the columns of Tables 2-5. Taking this average as a representative measure of reliability for a given segment for each method of calculation used, ranks from 1 to 6, according to size, were given the coefficients for the six segments listed in Tables 2-4 and ranks from 1 to 4 to the four corresponding segmental correlations presented in Table 5. The consistency of ranks for coefficients obtained by the four methods is taken as an indicator of the relative reliability of the different segments under consideration. Table 6 presents the array of these rank values from which this comparison may be made.

If one considers the rankings for the various segments, it appears that the one containing the whole series of trials (1-30) has the highest rating as determined by the four methods taken

TABLE 6

RANKS OF THE AVERAGES OF THE CORRELATIONS FOR SEGMENTS GIVEN IN TABLES 2, 3, 4, AND 5

Method of calculating r	Segment of Trial Series					
	1-30	1-20	11-30	1-10	11-20	21-30
A	2	5	1	6	3.5	3.5
B	1.5	1.5	3.5	3.5	4.5	4.5
C	1	3	2	6	5.5	5.5
D	3	2			1	4

collectively. Of the two twenty-trial segments, trials 11-30 would seem to have a slight advantage over trials 1-20 for the first three methods because of the low ranking given the latter by Method A (data for Method D not available for segment 11-30). In this connection, however, two important points should be noted. Firstly, the relatively low ranking of trials 1-20 given by Method A results from the extremely unstable performances of the animals during the first 3 to 5 trials when they are becoming adjusted to the maze situation. With them eliminated, the reliability as determined by this method is approximately equal to that of either of the other longer segments. Secondly, the rank of 2 given segment 1-20 for method D indicates that in this series of trials the animals are relatively consistent in the levels of their performances. This fact does not hold equally well for any segment having in it trials 26-30, in which most of the learning has already been accomplished. Furthermore, cutting off a few of the first erratic trials likewise improves the rating of segment 1-20 as determined by Method D. Hence, in view of the foregoing considerations, we believe that segment 1-20 is slightly preferable to segment 11-30, although the difference is slight and the choice of segment might be left to other considerations pertaining to the validity of the scores from these two segments. Probably the optimum segment of 20 trials is to be had by discarding trials 1-5 and 26-30 from the total series and retaining trials 6-25. Cutting off the last five is especially favorable to Method D; and the segment thus obtained most validly portrays the course of error elimination.

If we now compare the segments containing ten trials each, it is seen that the lowest ranking is given trials 1-10. Segments 11-20 and 21-30 are approximately equal in rank for Methods A, B, and C. For Method D, however, segment 11-20 is very much better. All things considered it would seem that the preferable segment of ten trials herein considered is 11-20. This

segment should also have high validity since it is free from the erratic performances marking the early trials and the lack of differentiation which occurs in the final stages when almost all erroneous responses have been eliminated.

The Stability of Reliability Coefficients in Relation to Animal Population

In this part of the study we have attempted to determine the relative reliabilities of coefficients of reliability as the population is increased by successive additions of groups of approximately 25 animals. To do this the data from which the coefficients presented in Tables 2-5 were computed were added in a cumulative manner, one group at a time, and in a jumbled order as far as the ages of the groups are concerned. Thus the data for the 8-months and the 5-months groups were added first; next the data for the 6-months group were added to the sum of the first two; then to the sum of the first three, the data of the 12 (a)-months group were added, etc. With each addition reliability coefficients were calculated anew. Since the data for the groups do not differ markedly with respect to their measures of central tendency and variability, no serious objection to this method of pooling is apparent. Fig. 2 gives the mean error curves for the groups for which data are pooled in this phase of the study. Inspection of the graphs reveals their close similarity of trend and shows that the groups did not differ greatly in their performances on the maze. Additional assurance of the legitimacy of this additive method is afforded by the consistency in magnitude of the coefficients of reliability within each column of Tables 2, 3, 4, and 5.

The coefficients of reliability obtained by this cumulative method of increasing the population and as calculated by Methods A, B, C, and D respectively are presented in Tables 7, 8, 9, and 10. A comparison of the coefficients obtained from the small groups as given in Tables 2-5 with the corresponding coefficients of the pooled groups of Tables 7-10 shows that the coefficients of reliability for small groups of twenty-five animals compare quite favorably with those obtained from much larger groups.[6] The greatest discrepancies occur in connection with segments that include the extremes

[6] In these comparisons it is to be observed that the population of the small group is included in the population of the large group. This may introduce a small factor making for correspondence between the two.

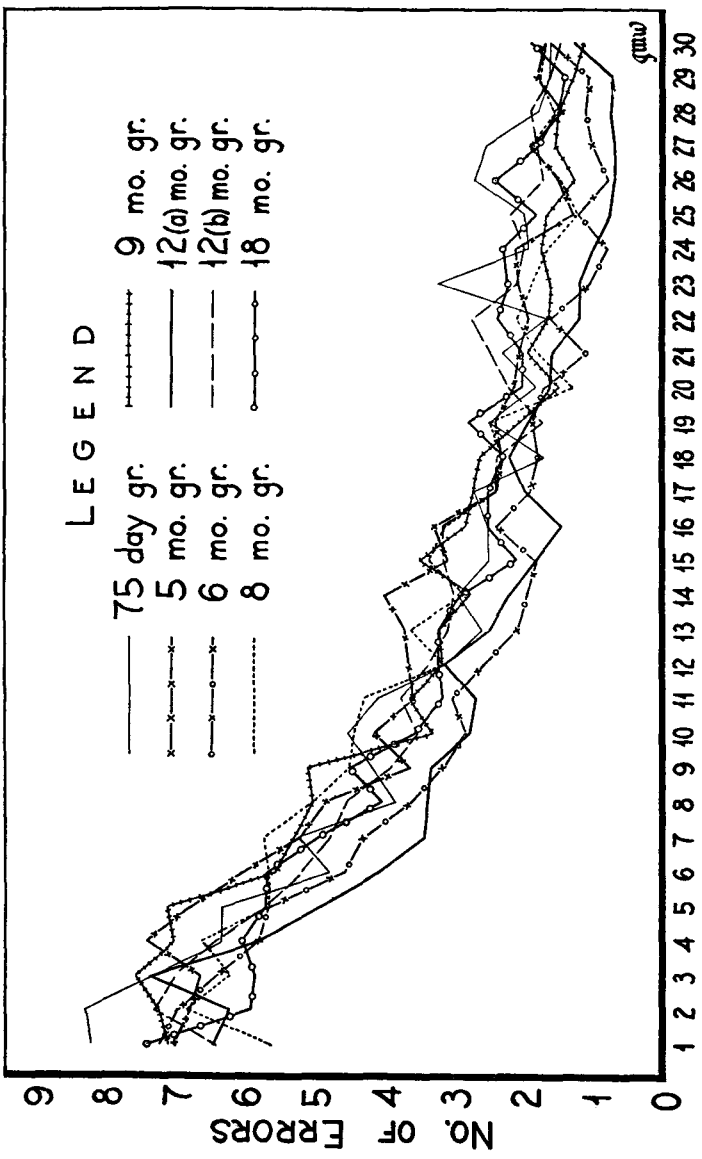

FIGURE 2.
MEAN ERROR CURVES FOR THE GROUPS OF ANIMALS CONSIDERED IN THIS STUDY AND LISTED IN TABLE 1.

of the trial series. These, as we have pointed out, are the least reliable portions of the total series.

For our present purposes of comparison it seems unnecessary to determine statistically the significance of the differences between coefficients of Tables 2-5 and Tables 7-10 by computing the ratio of each difference to the standard error of this difference. It will be of interest to note, however, the instances in which the probable errors for the individual groups listed in Tables 2-5 differ as much as three P. E.'s from the corresponding segmental coefficients based on the pooled data for all the groups. The significance of these differences as statistically determined will be less than that revealed by our simpler comparison. The results of tabulating these instances may be briefly stated as follows: there is about one chance in six that a coefficient appearing in Tables 2 and 3 will differ by 3 or more P. E.'s from the corresponding coefficient for the largest available population resulting from combining the individual groups; and there is approximately one chance in three that the coefficients in Tables 4 and 5 will differ from corresponding coefficients from the pooled data of Tables 9 and 10 by 3 or more P. E.'s.[7] As we have already stated, most of these discrepancies are associated with the unreliable portions of the trial series.

If we compare the correlation coefficients obtained from groups of 70 animals and 95 animals with coefficients obtained from the larger groups, we find that, with but rare exceptions, the coefficients of the former groups are all within 3 P. E.'s of the latter, irrespective of the method of calculating the reliability. This is readily seen by an inspection of the cumulative Tables 7-10. Another way of form-

[7]The coefficients computed from the group having a population of 16 have been omitted in this comparison as the number of cases is too small for the coefficients to be equally weighted with those of the other groups. In small groups there is a greater probability of the occurrence of high coefficients than is the case in larger populations.

The following equation gives the distribution of observed values of r when the true value of r is zero for varying values of n.

$$y = \frac{1}{\sqrt{\pi}} \frac{\left|\frac{n-1}{2}\right.}{\left|\frac{n-2}{2}\right.} (1-r^2)^{\frac{n-4}{2}}$$

If the above equation is plotted for small values of n, e.g., from three to ten, the truth of the above statement is readily seen.

TABLE 7

COEFFICIENTS OF RELIABILITY COMPUTED CUMULATIVELY BY METHOD A (ODD TRIALS VS. EVEN TRIALS)

Groups were added cumulatively in the following order: 8 mos., 5 mos., 6 mos., 12 (a) mos., 9 mos., 18 mos., 75 days, and 12 (b) mos.

	Segments of Trials					
N	1-30	1-20	11-30	1-10	11-20	21-30
25	.964±.005	.936±.008	.969±.005	.780±.033	.942±.008	.895±.012
41	.974±.002	.958±.004	.974±.002	.824±.021	.964±.004	.942±.007
70	.974±.002	.925±.007	.974±.002	.817±.016	.958±.003	.936±.005
95	.974±.001	.930±.005	.974±.001	.824±.014	.953±.003	.919±.006
120	.969±.002	.925±.005	.969±.002	.810±.014	.947±.004	.936±.004
145	.958±.002	.913±.005	.974±.001	.773±.013	.947±.003	.936±.003
170	.947±.003	.901±.005	.958±.002	.802±.011	.936±.003	.936±.003
205	.947±.003	.895±.005	.958±.002	.780±.012	.936±.003	.936±.003

TABLE 8

COEFFICIENTS OF RELIABILITY COMPUTED CUMULATIVELY BY METHOD B (ODD BLINDS VS. EVEN BLINDS)

Groups were added cumulatively in the following order: 8 mos., 5 mos., 6 mos., 12 (a) mos., and 9 mos.

	Segment of Trials					
N	1-30	1-20	11-30	1-10	11-20	21-30
25	.870±.020	.901±.012	.780±.049	.857±.020	.788±.033	.710±.047
41	.919±.009	.913±.009	.876±.014	.895±.010	.837±.018	.851±.016
70	.930±.007	.919±.007	.895±.008	.864±.012	.864±.012	.857±.012
95	.925±.006	.901±.007	.883±.009	.844±.012	.851±.010	.844±.012
120	.925±.005	.907±.006	.876±.008	.857±.009	.844±.011	.837±.011

TABLE 9

COEFFICIENTS OF RELIABILITY COMPUTED CUMULATIVELY BY METHOD C (FIRST ONE HALF OF BLINDS VS. SECOND ONE HALF OF BLINDS)

Groups were added cumulatively in the following order: 8 mos., 5 mos., 6 mos., 12 (a) mos., 9 mos., and 18 mos.

	Segments of Trials					
N	1-30	1-20	11-30	1-10	11-20	21-30
25	.925±.011	.844±.023	.844±.023	.773±.033	.788±.033	.693±.048
41	.901±.010	.901±.010	.889±.017	.795±.023	.876±.014	.788±.033
70	.925±.007	.901±.008	.907±.008	.817±.016	.870±.012	.837±.014
95	.913±.006	.889±.009	.895±.007	.851±.010	.844±.012	.837±.012
120	.895±.006	.870±.009	.876±.008	.817±.012	.824±.012	.817±.012
145	.901±.006	.864±.008	.889±.007	.802±.012	.830±.011	.851±.008

TABLE 10

COEFFICIENTS OF RELIABILITY COMPUTED COMULATIVELY BY METHOD D (SEGMENT VS. SEGMENT)

Groups were added cumulatively in the following order: 8 mos., 5 mos., 6 mos., 12 (a) mos., 9 mos., and 18 mos.

	Segments of Trials					
N	1-15 16-30	1-10 11-20	1-5 16-20	6-15 16-25	11-15 16-20	21-25 26-30
25	.765±.035	.810±.030	.462±.099	.947±.008	.876±.017	.870±.020
41	.864±.029	.864±.029	.667±.044	.942±.007	.913±.009	.930±.009
70	.864±.012	.883±.019	.693±.028	.919±.007	.907±.008	.942±.005
95	.795±.015	.851±.011	.649±.030	.895±.007	.895±.007	.870±.009
120	.824±.012	.857±.009	.639±.027	.895±.006	.901±.006	.817±.012
145	.788±.013	.857±.008	.621±.026	.876±.007	.895±.006	.795±.012

ulating this statement is to say that *beginning with the group of 70 animals the coefficients within each segment of the trial series assume values which undergo little change with the added increment of fifty or more animals to the population.*

RELIABILITY OF TIME SCORES

Since time scores are ofttimes used to measure learning ability on the maze, it seemed desirable to obtain reliability coefficients computed from them to compare with the coefficients obtained from errors. At this time it was not possible to calculate coefficients for all the groups listed in Table 1; hence two representative groups (75-day and 6-months groups) were chosen for consideration. The nature of our data permits the use of only Methods A and D to calculate these coefficients.

Table 2 gives the coefficients computed by Method A for various segments of the trial series. From an inspection of the coefficients it is apparent that the coefficients are uniformly high for all the segments and that, with a few exceptions, they compare favorably with the corresponding coefficients for errors. Taken as a group, the shorter segments embracing 10 trials yield coefficients that are slightly lower than those from the longer segments; in this respect the time scores are again in harmony with the error scores. From the standpoint of usefulness for experimental purposes the reliability of either short or long segments is sufficiently high for these time scores to warrant their being employed as measures of learning, providing their use can be justified on other grounds. Tolman and Nyswander (1927) regard them as less valid measures of learning ability than error scores.

Coefficients as calculated by Method D are likewise fairly high for the segments with which we have dealt. Table 12 contains typical coefficients calculated from the data of the 6-months group. Again these coefficients compare favorably with the corresponding coefficients for errors given in Table 5 and, like the latter, show that the animals tend to hold their relative places as to speed of running through the maze during short and long series of trials.

TABLE 11

COEFFICIENTS OF RELIABILITY FOR TIME SCORES COMPUTED BY METHOD A (ODD TRIALS VS. EVEN TRIALS)

		Segment of Trials					
Group	N	1-30	1-20	11-30	1-10	11-20	21-30
6 months	29	.947±.013	.936±.016	.969±.008	.913±.021	.969±.008	.870±.032
8 months	25	.864±.037	.824±.048	.942±.016	.758±.065	.876±.033	.895±.028

TABLE 12

COEFFICIENTS OF RELIABILITY FOR TIME SCORES COMPUTED BY METHOD D (SEGMENT VS. SEGMENT)

		Segment of Trials		
Group	N	6-10 vs. 11-15	11-15 vs. 16-20	6-15 vs. 16-25
6 months	29	.74±.064	.87±.032	.88±.030

RELATION OF OUR DATA TO THOSE OF EARLIER INVESTIGATIONS

From a consideration of the literature on reliability of animal learning scores it would seem to us that the four methods of computing reliability herein described are among the simplest and most direct methods applicable to the data as they are most frequently collected. We make this statement with strong reservations, however, because we have not yet had an opportunity to study at first hand the merits of other methods described in the literature. Furthermore, we strongly urge that other methods as well as the foregoing be applied to the same data in order that their individual and relative merits may be established more equitably than is now possible when they are applied to data collected on different instruments and by different investigators.

In a general way our data are consistent with the pioneer findings of Bagg (1920) who measured reliability of time

scores by a method that corresponds to Method D of our study. For the time scores of 93 mice running 17 consecutive trials on a relatively simple maze, he found a correlation of 0.46 for the sums of the time on trials 3-7 with the sums of the time for trials 13-17. (Trials 1 and 2 were discarded.) For similar five trial segments separated by five trials, the sums of time scores on the Multiple-T maze correlate to the extent of 0.75 to .95 for groups of 25 animals. We have nothing corresponding to his so-called "interference" test, which is essentially an instance of running animals on two different mazes and correlating the scores made on the two mazes. His "interference" test involved the use of the original maze in which certain alterations of doorways made each turn the opposite of that of the original. He obtained a correlation of 0.55 between the sums of the time scores for trials 3-17 on the first maze and trials 1-2 on the "interference" maze and a correlation of 0.49 between time scores for trials 3-17 on the first and trials 3-12 on the second.

An important variation of the second method of Bagg is that of Heron (1924) in which he attempted to ascertain the reliability of five stylus mazes with human subjects by intercorrelating the scores made on the several mazes. This method would give the reliability of either of two sets of maze data only if each maze measured exactly the same function and in exactly the same way.[8] If this is not the case, as probably happens when we consider mazes of different patterns and grades of difficulty, ambiguities in interpreting the results necessarily follow. Perhaps marked differences in the reliability of Heron's three criteria of learning (time, trials, errors) for the individual mazes accounts in part for the differences in the intercorrelations of these criteria in the two situations

[8] Since $r_{\infty_\omega} = \frac{r_{12}}{\sqrt{r_{11} r_{2II}}}$ (Kelley (1923), page 208)

then $r_{12} = r_{\infty_\omega} \sqrt{r_{11} r_{2II}}$

but r_{∞_ω} is never greater than 1

therefore $r_{12} \leq \sqrt{r_{11} r_{2II}}$

This expression which shows the dependence of the correlation between any two variables upon their reliability coefficients is sufficient to explain many of the low inter-correlations between mazes which have been used in experimental work.

presented by him, namely: (1) the correlations between scores of two different mazes (*with two different reliabilities*) and (2) the correlations between original scores and relearning scores on the same maze (*which has a constant or nearly constant reliability*).

Fundamentally Method A of our study is similar to the method used by Heron (1922) on a problem box in which he correlated the sums of time scores on odd with the sums of time scores on even trials; but there is one very important difference: namely, *the number of trials summated.* Heron required his animals to satisfy a criterion of accuracy before stopping their trials. As a result, there is a great variability in the number of trials required by the individual animals to satisfy this criterion of learning (range is 10 to 70 trials). Hence, even when the animals' successive performances are relatively inconsistent or when the inconsistencies for odd and even trials do not balance each other, as they tend to do in a fairly long series of trials, relatively high correlations between the sums of time scores on odd and even trials are to be expected because of the variations in the number of trials summated. If we understand his method correctly, it can yield only coefficients that are spuriously high *with a moderately reliable instrument and technique* but coefficients that are ambiguous and equivocal for interpretation and use unless some correction is made for the varying numbers of trials summated. Furthermore, it is important to note that discarding some of the early trial scores from the series, which can be amply justified in most cases, would probably tend in this instance to increase the divergence between the scores for different animals, since cutting off first trials affects the sums of time scores for the animals with short trial series relatively more than those of animals with long trial series. This fact may be partially responsible for his rather consistent rise in coefficients with successive dropping of early trials (up to the eighth trial). So far as we can see, Heron's data are not comparable with coefficients obtained from data in which the number of trials is constant for all animals.

The foregoing criticism appears to apply equally well to Hunter's data (1922) on the reliability of mazes based on the correlation of time scores of one segment of the trial series with another segment of the series. In this, as in the foregoing case, the number of trials run by different animals varied considerably since each was required to satisfy an arbitrarily de-

termined standard of accuracy and required different numbers of trials to do it. Hunter divides the trial series of each animal into tenths. These tenths were correlated to determine their inter-relationships without correcting for the difference in numbers of trials entering into each tenth. Hence the coefficients obtained are higher than one might expect from the same animal scores in which the number of trials divided into tenths was identical. From the data at hand Hunter obtained coefficients for adjacent tenths varying from 0.45 to 0.56 for a single unit T-maze, from 0.11 to 0.40 for a circular maze of intermediate difficulty, and from 0.31 to 0.69 from a circular maze of greater complexity. The coefficients were smaller as a rule for segments separated by one or more tenths.

We have no data comparable to those of Heron (1922) for a problem box, and Hunter and Randolph (1924) and Tolman and Davis (1924) for the maze in which they correlate the scores made on a short preliminary series of trials with a similar short series made after a period of inactivity. In certain respects the method of Bagg (see above) and Method D of our study resemble their methods insofar as we are dealing with correlations between segments of a trial series on the same instrument. Our methods differ, however, with respect to the length of interval between the two trial series and the activity of the animals during this interval. When, for instance, we correlate trials 1-10 with 11-20 only one day of inactivity falls between the two periods of learning and this is exactly the same as that falling between any of the trials of either series, whereas Heron, Hunter, and Tolman allow an interval of inactivity of 60, 30, and 6 days respectively. Bagg allows five days in which the learning tests proceed as before between his segments correlated. Neither Heron, Hunter, or Tolman obtained high coefficients of correlation by their method.

Reviewing Hubbert's data (1914) and relating it to her discussion of the merits of distance and time scores brings into question the reliability of the maze which she used. Three of her rats ran respectively 9,180 cm., 12,569.6 cm., and 30,488 cm., but learned the maze in the same number of trials. One speculates as to the character of the learning curves of these individual rats and the degree to which distance scores measured the differences between the animals. A method of obtaining a reliability coefficient which would cast some light on this question would involve obtaining partial correlation coefficients between the scores for distance on odd trials and similar scores

on even trials keeping the number of trials to learn constant. Such an index would give an idea as to the manner in which the curves of individual animals approached the same point on the base line, and hence to the validity of the use of distance as a criterion to distinguish between their performances. Such an analysis, of course, could be made with other criteria.

Other methods employing three variables occur which might be used to determine reliability. To illustrate, for Methods A and D which have been described in this paper, errors might have been correlated against errors with time constant for the odds and evens respectively or for the segments respectively.[9] Such correlations although they involve much labor would undoubtedly be of value in studies in which a more detailed analysis of the factors entering into maze behavior is desired. The combinations of errors, distance, time, retracings and perfect runs afford interesting combinations and their reliability may be determined statistically as indicated.

Conclusions

1. Coefficients of reliability for error scores on the Multiple-T maze are consistently high for short and long segments of the trial series. As computed by four methods, their ranges for groups of approximately 25 animals are as follows:

a. Method A—sums of errors on odd trials vs. sums of errors on even trials—r ranges from .59±.067 to .97±.004.

b. Method B—sums of errors on odd numbered blinds vs. sums of errors on even numbered blinds—r ranges from .71±.047 to .96±.007.

c. Method C—sums of errors for first half of the maze vs. sums of errors for second half of the maze —r ranges from .66±.055 to .96±.005.

d. Method D—sums of total errors for any segment of the trial series vs. sums of total errors for any other segment of the trial series—r ranges from .46±.099 to .97±.005.

---

[9] Let X_1=score in errors for sum of odd trials.
Let X_2=score in errors for sum of even trials.
Let Y_1=score in time for sum of odd trials.
Let Y_2=score in time for sum of even trials.

$$\text{Then } r_{(x_1.y_1)(x_2.y_2)} = \frac{r_{x_1x_2} - r_{x_1y_2} r_{x_2y_2} - r_{x_1y_1} r_{y_1x_2} + r_{x_1y_1} r_{y_1y_2} r_{x_2y_2}}{\sqrt{1-r^2_{x_1y_1}}\sqrt{1-r^2_{x_2y_2}}}$$

2. Coefficients calculated by Method A are slightly higher than those obtained by either of the other methods; those calculated by Method D, with a few exceptions, are slightly lower than those derived from either of the others; and those from Methods B and C occupy an intermediate position. All are sufficiently high to justify the use of these error scores, *so far as their reliability is concerned,* for experimental purposes.

3. With a few exceptions, reliability coefficients for our data as calculated by either of the four methods for groups of about 25 animals give close approximations of the coefficients to be obtained from groups of from 70 to 205 individuals. With rare exceptions those obtained from groups of 70 to 95 animals are within 3 P. E.'s of the coefficients obtained from groups of 205, irrespective of the method of calculating coefficients.

4. On the whole, the entire trial series gives slightly higher coefficients as calculated by Methods A, B, and C than any segment of the series. As calculated by Method D, however, coefficients based on the entire trial series are lower than those derived from segments that exclude the last five trials in which the learning curve almost parallels the base line.

5. The shorter segments of the trial series (10 trials) are less reliable than longer segments made up of 20 trials for each of the four methods of calculating reliability. One exception is noteworthy however; coefficients calculated by Method D for trials 11-20 rank well up with those for the twenty trial segments.

6. It is suggested that an optimum segment of the total trial series would probably be obtained by discarding the first two or three and the last five trials. This casting out of unreliable trials would be especially favorable to Method D and would not be very unfavorable to either of the others. By this method the validity of the error scores would probably be enhanced as well as the reliability of the data increased.

7. Coefficients of reliability for time scores from representative groups of approximately 25 animals as calculated by Methods A and D compare favorably with the corresponding coefficients for errors. As calculated by Method A, they range from .76±.065 to .97±.008 for trial series varying from 10 to 30 trials. As derived from Method D, they vary from .74±.064 to .88±.030 for segments of 10 to 20 trials.

8. In view of theoretical and practical considerations it

would seem expedient at the present time to determine the reliability of learning scores from mazes designed for the study of individual differences and similarities by at least two or more methods; each method carries with it some special significance. When knowledge of the consistency with which animals hold their respective places through the trial series is desired, Method D is especially recommended.

BIBLIOGRAPHY

1. BAGG, H. J. Individual differences and family resemblances in animal behavior. *Archives of Psych.*, 1920, **26**, No. 43.
2. BURLINGAME, EDITH MILDRED. Family resemblance in maze learning ability of albino rats. 1927 (M. A. Thesis, deposited in the Stanford University library).
3. HERON, W. T. The reliability of the inclined plane problem box as a measure of learning ability of the white rat. *Comp. Psych. Mon.*, 1922, **1**, No. 1.
4. HERON, W. T. Individual differences in ability versus chance in the learning of the stylus maze. *Comp. Psych. Mon.*, 1924, **2**, No. 8.
5. HUBBERT, H. B. Time versus distance in learning. *J. Animal Behavior*, 1914, **4**, 60-69.
6. HUNTER, W. S. Correlation studies with the maze in rats and humans. *Comp. Psych. Mon.*, 1922, **1**, No. 1.
7. HUNTER, W. S. AND RANDOLPH, VANCE. Further studies on the reliability of the maze with rats and humans. *J. Comp. Psych.*, 1924, **4**, 431-442.
8. KELLEY, T. L. Statistical method. New York: Macmillan, 1923.
9. KELLEY, T. L. Note on the reliability of a test; a reply to Dr. Crum's criticism. *J. of Educ. Psychol.*, 1926, **15**, 193-204.
10. MAUPIN, OAKLAND. Habit formation in animals. *Psych. Bull.*, 1921, **18**, 573-620.
11. SHEN, EUGENE. The standard error of certain estimated coefficients of correlation. *J. Educ. Psychol.*, 1924, **15**, 462-465.
12. SMALL, W. S. An experimental study of the mental processes of the rat. *Amer. J. of Psych.*, 1899, **11**, 133.
13. TOLMAN, E. C. AND DAVIS, F. C. A note on the correlation between the two mazes. *J. Comp. Psych.*, 1924, **4**, 125-136.
14. TOLMAN, E. C. The inheritance of maze ability in rats. *J. Comp. Psych.*, 1924, **4**, 1-18.
15. TOLMAN, E. C. AND NYSWANDER, D. B. The reliability and validity of maze-measures for rats. *J. Comp. Psych.*, 1927, **7**, No. 5.
16. WANG, G. H. Age and sex difference in the daily food-intake of the albino rat. *Amer. Jr. Physiol.*, 1925, **71**, 729-736.
17. WARDEN, C. J. The value of a preliminary period of feeding in the problem box. *J. Comp. Psych.*, 1925, **73**, 584-599.
18. WATSON, J. B. Animal education. 1903.
19. YOSHIOKA, JOSEPH. Weber's Law in the discrimination of maze distance by the white rat. *Psych. Bull.*, 1926, **23**, 295-297.

Stanford University *University of Utah*
California *Salt Lake City, Utah*