THE RELIABILITY AND VALIDITY OF MAZE-MEASURES FOR RATS

EDWARD C. TOLMAN
University of California
AND
DOROTHY B. NYSWANDER
University of Utah[1]

The following discussion and experiments arose from the attempt at an investigation of the inheritance of maze-learning ability in rats. (See the preliminary report on the latter by Tolman (15).) In the course of that investigation it soon became evident that a maze reliable for measuring individual differences must be discovered and also that some decision must be made as to the relative validities of the different possible maze-scores, i.e., number of blind-entrances, number of retracings, time, and number of perfect runs. The present article will discuss in detail these two needs. Part I will deal with reliability; and Part II with the relative validities of the different maze scores.

I. RELIABILITY

There has recently appeared in the literature much evidence for the unreliability of mazes as measures of individual ability in animals.[2] Reliability coefficients above 0.60 have seldom been obtained and the averages have been perhaps more nearly in the

[1] The writers wish to acknowledge their indebtedness to Professor C. P. Stone of Stanford University for helpful suggestions and for the use of data on his multiple T maze, and on the Carr maze; to Dr. Raymond H. Franzen for statistical advice; to Mr. R. C. Tryon for statistical computations and advice; and to Messrs. H. C. Blodgett, D. A. Macfarlane, O. L. Tinklepaugh for the use of their data.

The expenses of this investigation were met in part by grants from the Research Board of the University of California.

[2] See Hunter (5, 6, 7), Hunter and Randolf (8, 12), Heron (3, 4), Paterson (11), Liggett (10), Tolman (15), and Tolman and Davis (16).

neighborhood of 0.30. It is probable that if such figures had been obtained for mental tests that one was trying out for human subjects, such tests would have been rejected at once as not worth further bothering with. It seems, however, that with rats it is a question of either the maze or no method at all. For a maze seems to be the one instrument adapted to that animal's abilities, and if it proves valueless, then the chance of finding any other that will prove better seems hopeless indeed.

Given this situation, two things suggest themselves to be done: (1) We can, on the one hand, critically evaluate the meaning and significance of reliability coefficients to see what sorts of problems may still legitimately be attacked by the use of mazes even though the latter continue to give low reliability coefficients. And (2) we can, on the other hand, see if, by variations in maze pattern, technique of handling animals, and the like, we can not perhaps after all somewhat increase the reliability coefficients. We shall consider these two points successively in the two following sections.

The significance of reliability coefficients

A high reliability coefficient obtained with some given group of animals means that the given measuring instrument (maze) has tended to distribute all the individual animals within that group with great precision. It means, in short, that if these animals were measured a second time with the same instrument (without any carry over from the first experience) they would all tend to be placed in very much the same order with respect to one another the second time that they were the first time. If, then, one's problem is such that it requires that each and all the individuals within some group be accurately placed with respect to each other, a high reliability coefficient for the given measuring instrument will be essential. If, in other words, we wish to use mazes as measures of the differences between individuals *per se*, we must have high reliability coefficients.

It appears, however, that if, on the other hand, we are interested only in exhibiting either differences between individuals at the two extremes of the population, or differences between the mean performances of two very large groups, even though these mean performances lie relatively close together, such high reliability coefficients are by no means such a *sine qua non.* The common sense considerations which lead us to this conclusion can be seen from an analogy from simple physics. Consider, for example, a yard stick. If this yard stick be made out of elastic or some other material which expands and contracts to an enormous extent on the slightest provocation, or if the experimenter be very clumsy in handling it so that he induces great variations from moment to moment, or if, due to innate cussedness, the objects whose lengths are to be measured refuse to lie still so that the yard stick can really be precisely applied, the results for this yard stick are going to indicate a low reliability coefficient. Or, in other words, if the stick be used twice over for measuring a population of, say, logs varying from 1 to 2 feet in length, only a low positive correlation will probably be obtained between the lengths of all these logs as measured a first time and their lengths as measured a second time. The yard stick so applied results in no consistent distinctions between the true lengths of individual logs. And as far as the usableness of these resulting individual measures goes, it, of course, makes no difference whether their unreliability is primarily a function of something in the yard stick itself, of the clumsiness of the experimenter in handling it, or of the uncertainties and cussednesses of the logs themselves. But, it must be noted that notwithstanding this general unreliability of the yard stick for measuring individual logs *per se*, there would still be a high degree of probability in the case of two groups of logs really far apart on the scale, say a 1-foot group and a 2-foot group, that the *means* of these two groups would be successfully distinguished one from the other. And, secondly, it also appears that if we were interested merely in two *very large* groups of logs (some hundreds of thousands), even though they averaged nearer together on the scale, one set, say, ranging around an average of 24 inches, and the other around an average of 25 inches, then, if there were *large enough numbers* in the two groups, the chances would also be considerable that the average measure obtained for the 25-inch group would come out longer than that obtained for the 24-inch group. For in such a case all the merely chance variations due to the unreliability of the yard stick, the clumsiness of the experimenter, or the "jumpiness" of the logs themselves, would tend in the long run to cancel each other, and we should have left only the constant true difference between the averages of the two groups.

Or, to take a more familiar psychological example, it appears that even though the Army Alpha has a relatively low reliability coefficient within the group of college seniors, the chances are still great that, first, it would distinguish correctly between the average of a group of honor seniors and the average of a group of failing seniors; or, second, that if there were ten thousand seniors majoring in Psychology and ten thousand majoring in English, and even though Psychology-student intelligence be really *only a little* greater than English-student intelligence, it would distinguish correctly the direction of that difference. In other words, even though an instrument (as applied) is not precise enough, or consistent enough, to distinguish very reliably between all the separate individuals of a population (i.e., gives a low reliability coefficient), it may none the less be reliable enough to distinguish (*a*) between small groups at the two extremes, or (*b*) between the mean performances of very large groups even though the latter fall fairly near together on the scale. Two quantities falling very far apart or two quantities each measured a very very great number of times each have, even with an unreliable instrument, a good chance of being correctly distinguished.

Applying this to our mazes, it would appear, then, that, even though any given maze (together, of course, with its attendant method of application) be not reliable enough to place satisfactorily among themselves all the individual rats A, B, C, D, run under some one specific condition (for example, that, say, of one repetition a day), nor reliable enough to distinguish satisfactorily among themselves the individual rats a, b, c, d, etc., run under some other condition (for example, that of three repetitions a day), it may nevertheless be more than reliable enough to distinguish satisfactorily between the mean performances of these two groups of animals (if the two groups are large enough or if their mean performances happen really to fall relatively far apart upon the scale.

We conclude that even though the empirical evidence now at hand indicates that maze running gives low reliabilities for placing individual rats *per se*, this does not mean (as Hunter (7, 8) suggests) that all the work that has been done in the past, using such mazes to compare *mean* scores of large groups of rats, or groups of rats whose performances are really far apart, should therefore also be discarded. Instead, it would appear (See also Carr (2) for a substantiation of the present point of view) that all that is necessary is: first, that the obtained differences between the means of the contrasted groups shall satisfy the statistician's criterion of being, say, three times their own sigmas[3] (i.e., that differences of the given size shall be liable to occur by mere chance only some 27 times out of 10,000); and, second, that the numbers in the groups shall have been large enough to allow the application of statistical criteria.[4] For, as we have indicated, the pure logic of the situation indicates that valid differences of this sort can be obtained even with rats and mazes of fairly low reliability, provided only that the two values to be compared lie relatively far apart on the scale, or else that very great numbers of measurements be made.

The above argument, however, does not disprove the desirability of actually obtaining more reliable mazes, if such can be done. For, in the first place, even though we should wish to use

[3] It appears further from a recent statistical note by Tryon (17) that the usual formula for computing the sigma of a difference between two means should be corrected in terms of the reliability coefficients for the situations giving those means. And the correction is such that if these reliability coefficients are low, the obtained differences *become more rather than less significant*. In other words, provided only that the measuring instrument be sufficiently reliable to give a difference, then any such *obtained difference* will in reality be more rather than less significant, the greater the unreliability of the instrument with which it was obtained.

[4] Actually, it should be emphasized that much of the earlier work must undoubtedly be discarded also, not so much because the mazes used were unreliable, but because the numbers of animals in the groups were so small (5 or 10) that no statistical criteria of any sort, either as to maze reliability, or as to the significance of group differences, could properly be applied to them.

our mazes only for obtaining differences between group averages, obviously such differences are *more likely to show up* (with *reasonably small numbers* of animals) if the maze be reliable. And, in the second place, and this is important, it appears that to-day our purposes in animal psychology are such that we are often wanting to turn more and more to methods in which we should like to measure single individuals as such. We are turning, in short, to problems and techniques with animals analogous to those which the mental tester and the statistician have introduced into human psychology. On the one hand, we are becoming interested in individual differences *per se*, and, on the other, we are wanting to use individual differences and correlations as a technique for arriving at general relationships. Thus, for example, on the one hand we are now beginning to have investigations such as that on the inheritance of individual maze-learning ability, and, on the other, we are beginning to want to determine such general relationships as, for example, that between age and maze-learning ability; and in the doing of the latter it appears that it may well be helpful to do it not only by comparing the mean performances of large groups of animals at each of a few age-points, but also by running a large spread of animals at all sorts of age-points and determining a resulting correlation between age and ability. Further, it is to be noted that this latter, correlation, type of technique has a peculiar use in that it permits of the use of *partial* correlations, and thus the effect of certain unwanted secondary variables (which, perhaps, can be measured but not experimentally kept under control, such in this particular example as, say, weight) can be partialled out statistically.[5] It appears, in a word, that higher reliability coefficients for rat mazes, in spite of certain things that can be done without them, nevertheless are a consummation much to be desired. We return now, then, to the second of our two sub-problems: that of attempting *experimentally* to discover some shape of maze or techniques of running which shall, as such, succeed in giving a high reliability.

[5] It must be admitted, however, as has recently been beautifully demonstrated by Miss Barbara Burks, that the interpretation of partial correlations is subject to very grave pitfalls (1).

Reliability coefficient and shape of maze

We shall presently describe a series of mazes and the empirical reliabilities obtained from them. Before, however, we can be ready critically to evaluate these results. there are certain preliminary considerations to be noted.

Preliminary considerations

A. It is to be emphasized that the actually obtained reliability coefficients in any given experiment will be a function not only of constant and inevitable features about the maze itself, but also of the way in which any individual experimenter handles it. The technique of handling animals, control of environmental conditions, lights, noises, odors, control of internal conditions in the rat, such as hunger, etc., are factors which undoubtedly affect reliability, in addition to the mere structure of the mazes themselves. If these extraneous factors are carefully controlled, higher reliability coefficients are undoubtedly to be expected from one and the same maze pattern than if these factors are not so controlled. This, then, is the first caution to keep in mind in interpreting results. The actual data to be presented were obtained by a variety of different experimenters with presumably a variety of differences in the intimate techniques of handling animals, and the like. Hence it is probable that some of the apparent differences in reliability obtained were due not to the maze patterns as such, but rather to such differences in handling. It seems to us doubtful, however, that the major differences that we shall present as between mazes of different pattern are to be explained in this way.

B. A second caution to be considered in comparing our reliability coefficients has to do with the actual sigmas or true spreads of the particular groups of animals which happen to have been chosen. For it is evident that the size of any obtained reliability coefficient will depend in part upon the actual or true spread in abilities of the particular group of animals for whom that coefficient was obtained. If the group of rats chosen to test a given maze happened really to be very close together in true abilities, it is obvious that they will not be so likely to keep their respective places with reference to one another so consistently (i.e., a lower reliability coefficient will be obtained) than as if they were really much more widely separated in their true abilities. Thus, for example, the Army Alpha test on college students gives a much lower self-correlation than it does on a population with a wider actual spread in true abilities. But since, now, we have no way of telling to what extent the different groups of rats we happened to choose to test the various mazes may have been really equivalent in their true spreads (sigmas), it follows that some of the differences we obtained as between mazes may have been in part really a result of this extraneous factor of differences in spread between the different groups of animals that we ran. The chances again, however, seem strongly against the assumption that the major differences we obtained are so explainable. In all our cases each group of animals was chosen in a heterogeneous manner and included individuals from a number of different litters, so that it seems to us that the probability is that the true spreads in the different groups, while not, of course, equal, were nevertheless in most cases reasonably comparable.[6]

C. Finally, there is a third question which must be considered before we proceed to an examination of the actual data, the question, namely, of the method of computing reliability coefficients. For the situation in the case of a maze is relatively different from that of the ordinary mental test. In the mental test the standard procedure is to take the score on one-half of the *items* of the test and to correlate it with the score on the other half of the *items*. The coefficient of correlation between these two sets of half scores is said to measure the reliability of either half, and by a mathematical operation[7] it is possible to compute from it the probable greater reliability of the total scores which are to be obtained by adding the pairs of half scores together. And

[6] It must, however, be confessed that this matter of selection from different litters and control of the true spreads in the different groups was not given by us the careful attention it deserves. Professor Stone, in particular, has emphasized to the writers the importance of a more careful control of this matter of selection from different litters.

[7] That is, the so-called Brown-Spearman formula. Vide Kelley (9), p. 206.

an analogous procedure in the case of a maze would evidently be to take the error scores on one half of the blinds and correlate those against the scores on the other half of the blinds, provided that the behaviors with regard to both sets of blinds could be considered as equally good and independent samplings of the animals' ability. The different blinds would then correspond to the different items in a test. Actually, however, it would appear that the number of blinds in a maze is usually too few and their mutual interconnectedness too great for them to be divided into two chance halves in order that the scores on these halves may give anywhere near independent and equivalent samplings of the animals' ability. Such being the case, the usual procedure adopted seems to have become that of dividing not the maze but the learning curve into separate parts, and correlating the individuals' records on one of those parts of the curve against their records on another part of the curve. That is, error and time records on alternate runs have been correlated against each other (i.e., odd runs vs. even runs). Or errors and times on the first half of the runs have been correlated against errors and times on the second half of the runs. It is evident now that whichever of these procedures be adopted, the same fundamental assumption is involved. It is the assumption, namely, that maze-learning is, or can be made (by the use of a proper sort of maze), of such a nature that the stupid animal should consistently tend to exhibit more errors or time throughout all parts of learning than does the bright rat. If a rat tends to make more errors (take more time) than the average at the beginning of learning, he should, according to this assumption, tend to make more errors (take more time) than the average in the intermediate and last stages of learning also, and he should do it on both the odd and the even days. In other words, the assumption is that the individual learning curves should be parallel. Figure 1 represents a hypothetical case fulfilling such an assumption. Curve A is meant to represent a stupid rat who, after the first couple of runs (when all records will be equally subject to chance and hence are not to be counted), starts out poorly and makes more errors than the average for the beginning of learning, and who contrives consistently to make more errors than the average on all later runs as well. Curve B, on the other hand, is to represent an average rat, and curve C a bright one.[8]

But it may perhaps be contended that such an assumption is fundamentally wrong.[9] Thus, it may be pointed out that actually we often tend to obtain not parallel individual curves, but ones which might more nearly be represented by figure 2. Thus in figure 2, a rat, A, who (after the first couple of orienting runs, the results of which are to be discarded) starts out making more errors than the average, is depicted as none the less reaching the base line (i.e., beginning to make perfect runs) before the other rats C and B, who at the beginning (and, perhaps, in total) are

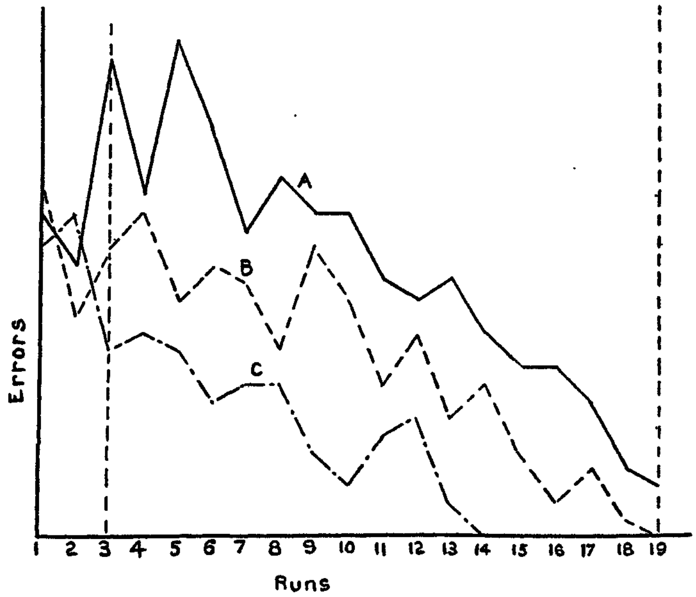


FIG. 1

[8] The two vertical dotted lines in the figure indicate that the data are considered only from trials 3 to 19 inclusive. The data up to trial 3 are disregarded because the rats up till then are all merely exploring in random fashion, and after trial 19 the data are discarded because from then on all rats are making perfect runs.

[9] The writers acknowledge their indebtedness to Professor Warner Brown for suggesting the possibility of this type of criticism.

depicted as making fewer errors. And it may be contended that such crisscrossing of curves as this is inevitable and an inherent feature of the maze-learning situation. It may be claimed, in short, that the shapes of the learning curves up to the point at which perfect runs are made are not significant, and that the only true measure of relative ability is simply this matter of how soon the curve reaches the base line (how soon perfect runs are made). Rat A, in figure 2, would, then, from this point of view be a good rat because his perfect run score (i.e., the total number of perfect runs made up until some constant number of total runs had been made by all rats) was high, irrespective of the sort of an error curve he had previous to this place where his curve hit the base line. Now such a contention is certainly a possible one. And we must consider its possible validity. For if it should prove valid, then our method of computing maze reliabilities by intercorrelating (error or time) scores on one part of the curve with scores on other parts of the curve must be given up. Further, it would appear also that no method of obtaining maze reliability

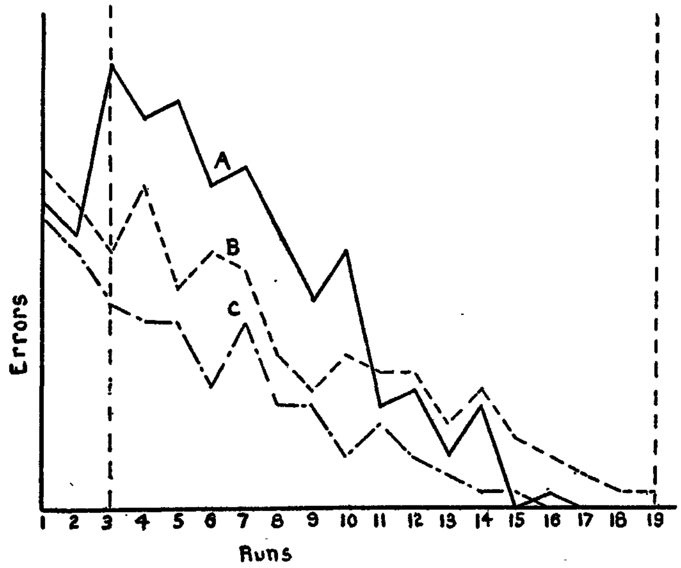


FIG. 2

coefficients from the records on a single maze would then be possible. For the perfect run score (or its inverse, the number of trials necessary to reach a given criterion of perfect runs), which would then become the sole criterion of learning, is obviously but a single entity. It is an item which can not apparently in any significant fashion be divided into independent halves. If we are to proceed, then, with the problem of reliability coefficients to be obtained from the single maze, we must first deal with such a possibility.

But what, now, are the arguments which are really implicit in and could alone support the contention that how soon the learning curve hits the base line is the important thing and that the various heights of the curve up to that point are not important? The only real argument for such a point of view which the writers can conceive of is to be found in the notion that a maze is for the animal something to be learned as a more or less closely interconnected whole. For, if such were the case, then the entering or not entering of individual blinds, up to the final point of learning, could be conceived as the "trying out of hypotheses," and not as, each in itself, a problem the reaction to which would be as such indicative. Only when the rat had finally got the whole idea, could he be expected to exhibit specific caution with respect to the individual blinds *per se*, so that his entering or not entering them would be, as such, indicative.

What, now, is the merit of such an hypothesis? Is a maze something to be learned as an interconnected whole where errors made or not made up to the final solution are not to be counted either for or against, or is it rather a thing which is learned piecemeal in a manner such that the successive elimination of errors does truly constitute a running commentary upon, or mirror image of the progress of learning? The answer is probably that some mazes are more of the one type and some more of the other. Thus it is conceivable that certain mazes which persist in giving crisscross curves are ones of the single interconnected whole type. For such mazes the only adequate measure of learning will be the perfect run score. And for them no measure of reliability, short of correlating this perfect run score with that of some second maze, will be feasible. It seems to the writers more probable, however, that most mazes belong to the other (or piecemeal) type, where the individual blinds are more or less independent units, so that learning is a matter of successive independent steps, and such that the successive heights of the error curve mark the successive progress in the taking of those steps. When, then, such a piecemeal maze is reliable, a rat, who (after the first trial or two, when he has had a chance to explore the whole maze) makes relatively more errors than the average, will do so because he is really stupid in acquiring the first steps to be learned. And he will also be stupid and make more errors than the average in acquiring all the later steps of learning. The different parts of learning in such a maze are to be conceived as relatively independent stabs at measuring the same thing. And in so far, then, as such a maze is reliable, it will tend to give separated, parallel curves like those in figure 1. And the intercorrelations of the error scores between successive parts of learning will tend to be high and to be the criterion of such reliability. When high correlations are obtained, it will mean that the given maze is a good one of the piecemeal variety. In other words, it would appear that a search for high intercorrelations between the different parts of the learning curve is a search for reliable *piecemeal mazes*. A maze which gives low intercorrelations may be either a maze of the total-pattern variety, or merely an unreliable piecemeal maze, but in either case it will be unsuitable for our purposes. If it is a maze of the total-pattern variety, we have no internal method of determining its reliability, and if it is merely an unreliable piecemeal maze, we do not want it.

One further question, however, must still be asked. Granting that it is a high correlation between the scores on all the different parts of the learning curve (i.e., a high parallelity of the individual curves) which we want, the question still remains as to what the two methods, that of correlating the scores on odd runs against those on even runs, and that of correlating the scores on the first half of learning against those on the second half of learning, respectively indicate. In order to answer this, let us again consider figures 1 and 2.

It will readily be perceived that in so far as the actually obtained data approximate very closely the perfect degree of parallelity represented by figure 1, either way of computing will give the same answer. For in so far as the individual curves approximate closely to actual parallelity, then the poor rats will be consistently different from the good rats on alternate runs, on even runs, on the first half of the runs, on the second half of the runs. And the correlations of odd vs. even and of first half vs. second half will both be high and approximately the same. Consider, however, a possibility such as that represented in figure 2. For such a case there would be quite a high correlation between scores on odd runs vs. even runs, but practically no correlation between scores on the first half of the runs vs. scores on the second half of the runs. This follows because the scores on the second half of the runs are represented as undifferentiated or possibly differentiated in some different way from those in the first half. Considering, however, the odd vs. even correlations, it appears that this, considering the first half of learning alone, would give practically a unity correlation. Further, it appears that if the scores in the second half of learning are undifferentiated, or even consistent among themselves in some new way, provided this new way is not too negatively correlated with the first part of the curve, the total odd vs. even correlations can still remain fairly high. Only, if the differentiation of individuals in the second half of learning were consistently the reverse of that in the first half of learning, would the odd vs. even correlation go down in the same way as the half vs. half correlation. We conclude, then, that when one half of the curve (either first or last) tends consistently to differentiate the individuals within itself and the rest of the curve does not differentiate them, or differentiates them in some way not highly correlated with this first way, we shall generally find a higher correlation of odd vs. even than of half vs. half. A high odd vs. even correlation means consistent differentiation of individual animals, averaging the curve as a whole, but an equally high half vs. half correlation means in addition that this consistent differentiation is due to the measurement of one and the same function *throughout* the curve. The best maze, then, will be one which, if possible, gives both high odd vs. even and high first half vs. second half correlations.

We may briefly recapitulate with regard to our three preliminary considerations:

A. Obtained reliabilities will presumably depend in part upon such factors as constancy in the technique of handling the animals, in controlling their hunger, in controlling general experimental conditions, etc., as well as upon the mere maze-patterns as such.

B. Obtained reliabilities will depend also upon the actual true spreads in ability of the groups of animals tested. If the animals in a group differ but slightly in real ability, a smaller reliability coefficient will be obtained on the same maze from them than would have been obtained from a group in which the animals had happened to lie further apart in real ability.

C. Reliabilities are to be evaluated in terms of both odd vs. even and first half vs. second half correlations. High half vs. half correlations, as well as high odd vs. even correlations, will indicate not only that the individual animals are consistently differentiated, taking the learning curve as a whole, but also that the two ends of the curve are tending to measure the same thing.

Results

Table 1 presents the data from fourteen experiments involving seven distinctly different maze patterns.

The first column gives the number of the experiment. All experiments with the same Roman numeral involved mazes of approximately the same pattern.

The second column indicates briefly the type of maze.

The third column designates the experimenter. N and T stand for the writers, Nyswander and Tolman, respectively; S stands for Professor C. P. Stone of Stanford; Ti stands for Mr. O. L. Tinklepaugh; B for Mr. H. C. Blodgett; Mac for Mr. D. A. Macfarlane; Sh for Miss Dorothy Sherman; Bu for Miss M. Burlingame. The latter was a graduate student at Stanford University. All the others, save the senior writer and Professor Stone, were graduate students at the University of California, except Miss Sherman who was a senior at California. All experiments, save III and VII, were performed at California. The latter two were performed at Stanford under Professor Stone's direction.

The fourth column indicates the number of animals in the group which was tested and from which the reliability coefficients were computed.

TABLE 1

Errors

EXPERIMENT NUMBER	TYPE OF MAZE	EXPERIMENTER	NUMBER OF ANIMALS	RUNS	ERRORS ODD VS. EVEN	ERRORS ½ VS. ½
I	Multiple choice, 12 blinds	N	23	3–10	0.482	0.035
II	Circular, 5 blinds	Ti	24	3–14	0.563	0.223
III	Carr	S, B	42	3–16	0.597	0.366
IV (a)	Simple 3-way, 2 blinds	N	20	3–25	0.604	0.453
IV (b)	Simple 3-way, 2 blinds	T	21	3–10	0.440	0.344
IV (c)	Simple 3-way, 2 blinds	N	18	3–10	0.447	0.595
IV (d)	Simple 3-way, 2 blinds	N	15	3–10	0.504	0.506
V (a)	Forward going right left, maze, 7 blinds	Ti	24	2–17	0.686	0.337
V (b)	Forward going right left maze, 9 blinds	Mac	18	3–32	0.880	0.589
V (c)	Forward going right left maze, 9 blinds, Wading	Mac	20	3–32	0.698	0.351
VI (a)	T maze, with doors, 6 blinds	B	36	3–10	0.587	0.600
VI (b)	T maze, without doors, 6 blinds	N	28	3–10	0.720	0.652
VI (c)	T maze, with doors, 8 blinds	Sh	26	3–10	0.559	0.415
VII	Multiple T maze, 3 doors, 14 blinds with dumbbell ends	Bu	25	4–19	0.833	0.821

The fifth column indicates the runs, the data from which were included in computing these correlations.

Columns six and seven present the correlations obtained between scores on the odd runs (or days) vs. scores on the even runs (or days), and between scores on the first half of the runs (days) and scores on the second half of the runs (days), respectively.

The immediately succeeding paragraphs in fine print indicate in more detail the nature of the individual maze patterns and the techniques.

Experiment I. Multiple choice maze. This was a maze consisting of four successive banks of alleys. Each bank was made up of four adjacent parallel alleys two feet deep. One alley in each of the four was the true path; the other three were blinds (12 blinds in all).

Technique. The technique of handling, etc., was probably somewhat above average. The animals, however, were not weighed to control food incentive. One trial per day.

Experiment II. Circular maze. This was a simple maze of the general Watson type (5 blinds). It was provided, however, in a way that the usual Watson maze is not, with doors (operating on the guillotine plan) over the passages from one concentric ring to the next. These were closed by the experimenter behind the animal so as to prevent retracings.

Technique. There are several points to emphasize. The animals run on this maze had been previously run on maze V (a), so that they were thoroughly used to being handled. Environmental noises and lights were more carefully controlled than is usually the case (a fan was kept going to equalize noise), and the animals were weighed regularly and the amount of their diet carefully controlled. And, lastly, they consisted of two groups of 12 each, one of which groups was being injected daily with saline solution, and the other of which was being injected with a specially prepared hypophysis extract. Hence the total group of 24, if there were any divergent effects as between the two types of injection, should have had a relatively large true spread or sigma. In a word, the technique was decidedly above the average and should, as such, have tended to give high reliability coefficients. One trial per day.

Experiment III. The Carr maze. This is a well known maze, and it, along with the average learning curve obtained, is diagrammed in figure 3.

Technique. Here also the technique was probably decidedly above the average. The environment was kept constant and the animals were weighed each day before running, and their diet regulated accordingly. In particular it is to be noted that the technique was the same as that in experiment VII (the case which gave the highest reliabilities of all). One trial per day.

Experiment IV (*a*). *Simple 3-way maze.* This was a maze in which the animal entered a choice box from the opposite side of which three immediately adjacent alleys opened. The middle of these three constituted the true path and led directly through a distance of 2 feet to food. The two alleys, one on either side, were blinds. These blinds ran along side of and adjacent to the true path for its full length. At the place at which the latter entered the food box, they then each made a right angle projecting off to the right and to the left. In this experiment one of the blinds after turning the corner projected out only one foot, the other projected out three feet.

Technique. The technique of handling, etc., was average. The rats had been run in experiment I. One trial per day.

Experiment IV (*b*). *Simple 3-way maze.* The maze was identical with that in experiment IV (a), save that the alleys were all 6 inches wide instead of only 4 inches wide.

Technique. Average. One trial per day.

Experiment IV (*c*). *Simple 3-way maze.* The maze pattern was the same as in experiment IV (a), save that both blinds were of equal length and projected 3 feet beyond their elbows.

Technique. This was a *relearning* experiment. The animals used were the same ones that had been previously run in IV (a). An interval of 63 days intervened between learning and relearning. The method of handling, etc., was average. One trial per day.

Experiment IV (*d*). *Simple 3-way maze.* The maze-pattern was the same as in experiment IV (c).

Technique. The animals were ones which had been previously used several weeks earlier by a different experimenter, B, in experiment VI (a). The method of handling, etc., was average. One trial per day.

Experiment V (*a*). *Forward going right left maze.* This maze, with its learning curve, is diagrammed in figure 4. It was devised by Mr. O. L. Tinklepaugh. The entrances from one alley to the next were provided with guillotine doors which were successively closed behind the animals to prevent retracing.

Technique. The animals were the same 24 which were subsequently run on maze II. And, as in the case of maze II, the technique of handling, etc., was superior. There were two runs a day, one in the morning and one in the afternoon. In computing reliability coefficients, each day was treated as a unit.

Experiment V (b). Forward going right left maze. The maze in this experiment was of the same general pattern as that in Experiment V (a). It had 9 choices instead of 7. The alleys were 5 inches wide instead of 4, and there was only one door about half way through the maze.

Technique. The alleys were filled with water so that the animals swam through instead of running through. There were no return runs, but at the exit there was a platform on which the animals could climb out of the water and be fed. The use of this water probably provided a particularly constant motive. After the first two days there were 3 runs a day, separated by approximately one half hour intervals. In computing reliability coefficients, each day was treated as a unit.

Experiment V (c). Forward going right left maze. The maze is the same as in experiment V (b).

Technique. The animals swam through the first four alleys and then waded through an inch of water the rest of the way. Otherwise the conditions were the same as in experiment V (b). Three trials per day.

Experiment VI (a). Six-unit T maze with doors. This maze was devised by Mr. H. C. Blodgett, and a diagram of it is shown in figure 5. It will be seen that the maze divides up into a succession of equivalent T units. Another feature was that there were doors at the end of each section of true path, which were closed behind the animals and prevented retracing. In this way uncontrolled differences of exercise in the different parts of the maze were prevented.

Technique. The technique was average. One trial per day.

Experiment VI (b). Six-unit T maze without doors. The maze-pattern is the same as that in experiment VI (a), save that the doors were omitted and retracings allowed. Blind entrances in retracing, as well as in forward going, were counted.

Technique. The technique was average. One trial per day.

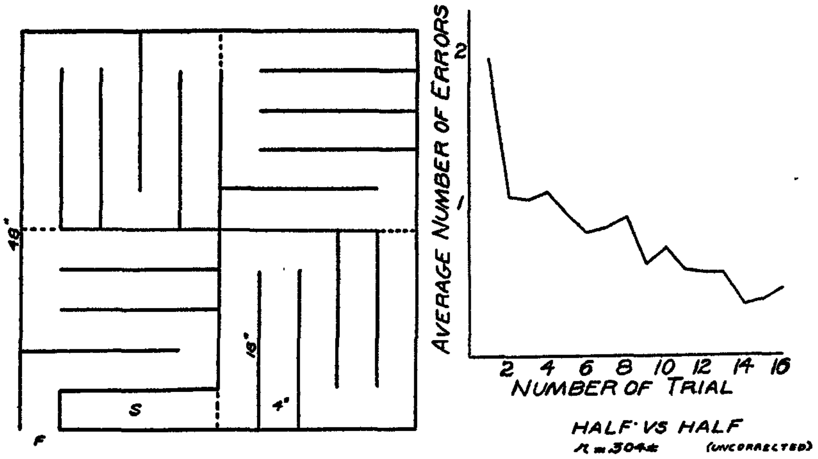
Experiment VI (c). Eight-unit T maze with doors. The maze-pattern was the same as in experiments VI (a) and VI (b), save that two more T units were added.

Technique. The technique was average. It was noticed, however, that for some reason a very small spread in scores was obtained. All the animals tended to make an unusually large number of perfect runs, and this probably tended to bring down the coefficients. One trial per day.

Experiment VII. Multiple T unit maze. This was devised by Professor Stone as an improvement on the Blodgett 6-unit maze. A diagram of it is shown in figure 6. In addition to the greater number of units, an outstanding feature of it is the "dumbell ends" at the end of each blind, thus making the blinds "look" like true paths. There were only 3 doors. These are indicated by the dotted lines.

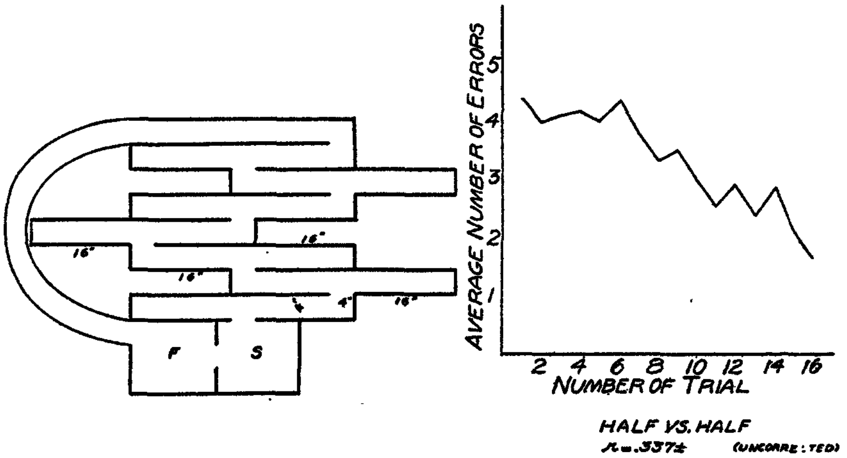
Technique. The technique was above average. The animals were weighed daily to control hunger, and the general conditions were good. One trial per day.

Considering now the results as shown in table 1, certain preliminary points are to be noted: (1) The table presents error records only. The reason for neglecting the time records will appear below in Part II under the discussion of validity. (2) The records of a certain number of the beginning trials have been excluded in each case. This is because the number of errors made upon the first trials (before the animals have had any previous experience of the maze) is necessarily a matter of chance, and the size of the error scores is so great in these first trials that, if included, they tend unduly to weight the total scores. (3) The records were also in each case all cut off beyond some given number of trials. The determining of this point of cutting-off was in each instance more or less empirical. Our general rule was to cut off at the point where a fair proportion of the animals had begun to give zero scores (i.e., to make perfect runs). In general, it is to be noted that whereas the inclusion of too many final zero (perfect run) scores would apparently have little effect upon the odd vs. even correlations, since the inclusion of them would merely add zeros to both sides of the pairs of figures to be correlated, the inclusion of too many of these zero scores would tend to lower the half vs. half correlations. For if there were too many of them, they would tend to erase the distinction between individuals in the latter halves of their scores.



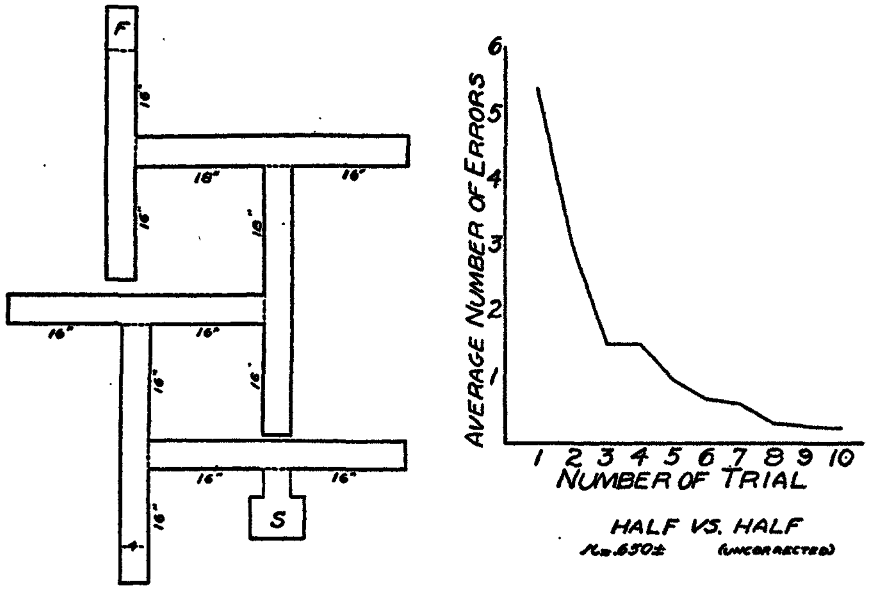
CARR MAZE LEARNING CURVE

FIG. 3



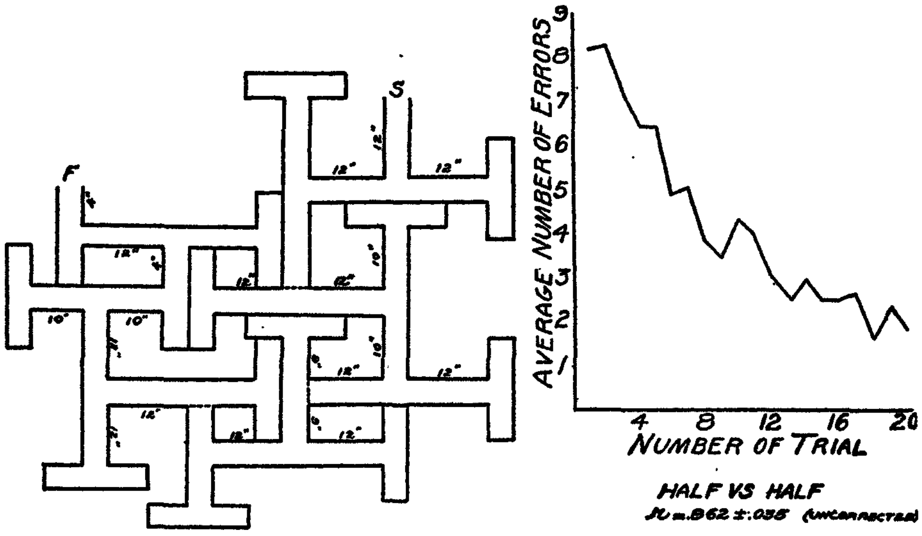
RIGHT-LEFT MAZE LEARNING CURVE

FIG. 4



T-MAZE LEARNING CURVE

FIG. 5



MULTIPLE-T MAZE LEARNING CURVE

FIG. 6

The ideal way would, of course, be to run all the rats well beyond the maximum number of trials to be retained and then, by experimentally cutting off the data one trial at a time, to have determined statistically just the number of trials which would have given the maximum correlations, both odd vs. even and first half vs. second half. Such a procedure should be adopted for any maze finally accepted as the one to be used for a long series of experiments. The large amount of labor, however, which would have been involved in the present instance did not seem justified, considering the wholly *preliminary* nature of this investigation.

Turning now to a comparison of the figures in the table, it will be noted that we have arranged the order of the experiments in such a way as to give in general an increasing half vs. half correlation as we proceed down the table. Thus the half vs. half correlation is lowest for experiment I (using the circular maze) and highest for experiment VII (using the multiple T maze). The experiments and mazes in between give half vs. half correlations intermediate in size. If we examine the resulting arrangement of the odd vs. even correlations, we note, first, that these latter in every case save one (IV (c), which is probably some sort of a fluke) are higher than their parallel half vs. half correlations. And, secondly, we note that they do not follow nearly so distinct a course from low to high as do the odd vs. even correlations. They tend much more nearly to be all of about the same size and relatively high. In other words, in terms of our preceding discussion as to the relative meanings of half vs. half and odd vs. even correlations, it would appear that while most of the mazes were fairly good at differentiating individuals, given the learning curves as wholes, all the mazes were not nearly so good at measuring one and the same thing at both ends of the curves. (This fact that learning curves do not measure the same thing throughout has also been demonstrated by Ruch for human beings (11, 12)).

Three typical cases where the odd vs. even correlations are noticeably higher than the corresponding half vs. half correlations are experiments II, V (a), and V (b). If, now, our hypothesis as to the meaning of these discrepancies is correct, then if in these cases we obtain reliability (odd vs. even) coefficients for each end of the curve by itself and then correct the intercorrelation between the two ends for these reliabilities, the so-called correction for attenuation (9, p. 209), the half vs. half correlations should still stay well below unity, indicating that the two ends of the curves were not completely measuring one and the same thing. If, on the other hand, we perform a similar operation for typical cases, such as experiment VI (b) and experiment VII, in which the half vs. half correlations come up to the odd vs. even correlations, here the half vs. half correlations, thus corrected for attenuation, should come up to unity. That these are the respective outcomes is shown in table 2.

TABLE 2

EXPERIMENT NUMBER	RUNS	OBTAINED ODD VS. EVEN	CORRECTED BY BROWN SP.	OBTAINED ½ VS. ½	½ VS. ½ CORRECTED FOR ATTENUATION
II	3–8	0.310	0.473	3–8 vs. 9–14	3–8 vs. 9–14
	9–14	0.705	0.827	0.223	0.357
V (a)	2–9	0.257	0.409	2–9 vs. 10–17	
	10–17	0.802	0.890	0.337	0.559
V (b)	3–14*	0.549	0.709	3–14 vs. 21–32	
	21–32*	0.768	0.869	0.565	0.720
VI (b)	3–6	0.632	0.775		
	7–10	0.285	0.444	0.652	1.111
VII	4–11	0.716	0.834		
	12–19	0.735	0.847	0.821	0.9768

* In this case we had such a long curve that we could leave out some of the middle and still have extreme ends of the curve each having a very considerable reliability within itself.

Conclusions and interpretation

Returning again, then, to table 1, we may conclude: (1) that maze types V, VI and VII, that is, the forward going right left and the T mazes, are decidedly good. And we may conclude

(2) that of the two types the T maze is the better, for it gives high half vs. half, as well as high odd vs. even correlations. (That is, it, the T maze, tends to measure one and the same thing at both ends of the learning curve in a way that the forward going right left maze does not.) We may conclude (3) that of the T mazes the multiple or 14-unit one is best of all. And, finally, we may conclude (4) that such differences of technique as existed between the different experiments appear to have been for the most part of minor importance as compared with the effect of gross differences in maze pattern.

How are these results to be interpreted? First, what is there common to both the forward going right left mazes and the T unit mazes which makes them both good? The answer is necessarily speculative. But it seems to the writers significant that in both these types of mazes the blinds are placed in such a way that each blind always presents a direct alternative with a complementary true path. And, secondly, we would also draw the reader's attention to the fact that the blinds and the true paths were all of the same lengths. The mazes, that is, divided up into sets of equal units. These two factors taken together constitute, we believe, the essential virtues of these mazes.

Secondly, we may speculate as to why the T mazes tended to give higher half vs. half correlations than did the right left maze. The answer, as the writers see it, is that in the later stages of the right left maze the animal as he acquires some speed of running tends to overshoot the mark. In passing out of one alley into the next, he tends to get carried into a blind simply because heis going too fast and does not check himself in time. A considerable number of such errors were actually observed by Mr. Macfarlane (experiments V (b) and V (c)) in the later stages of learning. But, such being the case, then it is evident why the later stages of learning in which the error scores tended to depend in part upon this factor did not tend altogether to measure the same function as that which was measured in the first part of learning. In the multiple T unit maze any such overshooting tendency was minimized by the stems of the T's which intervened between the successive choices.

Finally, we have the question as to why of the T mazes the 14-unit maze is the best of all. The answer would seem to be that having a greater number of units is equivalent to having more items in a test. It gives a longer learning curve. The more data thus obtained, the greater the minimization of chance in the final scores. (The fact that the 8-unit T maze, instead of being better, was worse than the 6-unit T maze, we would explain and reconcile with this conclusion by the fact that for some reason or other the particular group of animals run on the 8-unit maze happened to have a very slight spread in individual ability. They all began making perfect runs (i.e., obliterating their individual differences) at an unusually early stage in the learning curve). One reason why the earlier mazes presented in table 1, such as the Carr and the 5 blind circular maze, were particularly bad may well be because of the smallness of their numbers of blinds. Thus Warden (18) has shown by a detailed analysis of his data that only 4 out of the total 10 blinds really affect to any appreciable degree the total score in the Carr maze, and an inspection of the learning curve in figure 3 immediately shows the same thing.

So much for maze-pattern as the determiner of reliability. Mazes built up of successive units, where each unit is as nearly as possible equivalent to every other, would appear the best. And of these, T unit mazes would appear better than simple parallel unit mazes. Lastly, mazes containing great numbers of units would appear the best of all.[10]

Our final conclusion that differences of technique were relatively negligible in their influence as compared with that of the maze pattern is substantiated by a number of points. Thus, for example, it is to be noted that two among those which we evaluated as the best techniques of all (i.e., the technique of Professor Stone and his co-workers) gave, respectively, the highest and lowest reliabilities (vide experiments VII and II). Thus, it appeared that the common goodness of technique was more than

[10] It is to be noted that successive unit mazes would *a priori* appear to be ones most likely to cause a piecemeal variety of learning which, as we decided above, would necessarily be the type most likely to give us high intercorrelations.

snowed under by the pronounced divergence in the mazes. Again, examining the results for experiments IV (c) and IV (d), which were, respectively, a relearning experiment and an experiment using rats which had previously been trained in another maze, there appears no superiority in them as compared with the results of experiment IV (a), although it might have been expected that the previous handlings which the rats had had before they came to experiments IV (c) and IV (d) would have tended to increase the stability and consistencies of their daily performances and to give higher reliabilities than that for experiment IV (a) in which there had been no such previous handlings.

Only in two places is there any evidence that differences of technique did definitely affect the results. On the one hand, there is the fact that the 8-unit T maze gave poorer reliabilities than the 6-unit T maze, which (as we have already indicated) is, we think, to be explained by the small actual spread among the animals which happened to be run in the 8-unit maze. And, on the other hand, there is the fact which appears to come out from a comparison of experiments V (b) and V (c). In experiment V (b) the rats swam through the maze, while in experiment V (c) they ran through it. And it appears that experiment V (b) gave definitely higher reliabilities than experiment V (c). The explanation presumably is that it was the increased strength and consistency of the motive brought about by the condition of having to swim through which caused the improvement in reliability.

Summary and conclusions

We may now briefly sum together these more important results and conclusions as to reliability.

1. High reliability coefficients are a *sine qua non,* only if we wish to measure all the separate individuals of a population *per se.* If we wish merely to compare the means of *large* groups of individuals, or to compare single individuals lying at relatively *distant points* on the scale, measuring instruments (i.e., mazes) with relatively low reliability coefficients will be adequate.

2. The general usefulness of our mazes will nevertheless be greatly increased if we can evolve ones which will give high re-

liability coefficients. For we can then attack problems involving individual differences *per se* (such, for example, as the inheritance of ability), as well as problems of a more general nature in which it would be helpful to use the correlation technique.

3. As to the method of computing the reliability coefficient for a maze, it is evident that the theoretically best method would be to have long enough mazes and mazes of such a structure that the scores on individual blinds could be considered as separate, independent items. Then these could be separated into two independent halves and correlated one against the other. With the mazes as yet tried out, however, such a method has not been feasible. The mazes have been too short and their scores on the separate blinds have for the most part not been sufficiently independent of one another for such an attempt to having meaning.[11] In default of a correlation between scores on separate groups of blinds as an adequate measure of reliability, the remaining possibilities would appear to be either correlations between scores on odd runs vs. even runs, or correlations between scores on the first half of learning vs. scores on the second half of learning. And it appears, further, that whereas the correlation of odd vs. even indicates the consistency (reliability) of the individual scores, an equally large correlation of half vs. half is also needed to demonstrate that these individual scores are due to the measuring of one and the same thing throughout the whole length of the learning curve.

4. Our actual data as to different maze patterns indicate that a multiple unit type of maze is the best; that, of these, the T unit is better than the simple right left unit; and that 14 units are better than 6 or 8. (The 14-unit maze gives raw correlations between the separate parts of the curve of 0.833 and 0.821, which, when corrected by the Brown-Spearman formula to indicate the probable correlation of the total score against that from another

[11] It should be noted, however, that the multiple T unit mazes, which, on the basis of our method of correlating parts of the learning curve against one another, gave the highest reliability, would *a priori* appear to be approaching in form a type of maze which might be expected to give high correlations between scores on separate groups of blinds. Further work should be done on this matter.

similar curve, if the latter could be obtained, go up to 0.909 and 0.902, respectively.) In short, *mazes can be constructed which give high reliability coefficients even for rats.*

II. VALIDITY

Finally, we shall consider now under this head of validity the relative merits of time, retracing, and perfect run scores as compared with the error scores which we have alone considered thus far. And for the purposes of this argument we shall illustrate by experiment VI (b). This experiment used a good average maze, and we happen to have all the requisite data from it at hand.

Time

To any one who has run rats in mazes it is obvious that individual animals would seem to differ in something to be called pure speediness. And, further, whatever this pure speediness may be, whether it is to be conceived as a matter of mere physiological tempo, or of emotional or temperamental factors (i.e., timidity, cautiousness, etc.), it is probably something in no intimate way related to error-eliminating ability. Now, time scores depend in part upon this pure speediness, but in part also upon the fact that errors, as such, take time. Gross time scores must, then, be ambiguous since they are made up of these two variables which are not necessarily related,—"pure speediness" on the one hand, and "error-eliminating ability" on the other.

But there is still another reason why time is a bad, ambiguous, type of score. This will be brought out by concrete figures. Suppose, for example, that for a typical maze we intercorrelate the three measures: errors (i.e., blind entrances) (runs 3–10) E; time (runs 3–10) T; and number of perfect runs (runs 3–10) P. We obtain for the Blodgett maze experiment VI (b), table 1 ($n = 28$) the following figures:

$$r_{ET} = 0.606$$
$$r_{EP} = -0.838$$
$$r_{TP} = -0.332$$

The examination of these figures indicates, first, that, as was to have been expected, time correlated highly with errors (since, as we have just indicated, the time score will be in large part due to the number of errors). We also see that errors and perfect runs give an even higher negative correlation. (That is, the individual curves run relatively parallel and like those in figure 1.) The one new and unexpected feature is the lowness of the negative correlation obtained between time and perfect runs. It appears, in other words, that among the things in time which make it a different score from errors is something which makes it give a lower negative correlation with perfect runs than does errors. And the truth of this situation becomes even more striking if we work out the partial correlation. For if we compute the partial correlation between T and P with E constant, we obtain the following positive figure, $r_{\mathrm{TP.E}} = 0.406$. In other words, when errors have been made constant, it appears that time, as such, has a *positive* correlation with perfect runs. Those animals making the most perfect runs are also those taking the most time.

But what is the meaning of this? A first suggestion which presents itself is that since time, as such, is, as we indicated above, closely related to timidity or cautiousness, then within groups of animals all making the same numbers of errors (which is what we are comparing by the method of partial correlation), the slower animals are the more cautious ones, and hence they are also the ones who make the most final perfect runs. The ones with less time would, on the other hand, according to this view, be the incautious rats who, from mere haste and carelessness, continue to make a small number of final errors, and hence to lower their perfect run scores. Such an interpretation was indeed the one arrived at by Tolman (15) in the earlier study on inheritance already referred to and in which a similar finding was obtained.

It appears, however, that a more careful statistical analysis suggests a second, more probable, explanation. Consider figure 7. We have drawn the case for three animals, all of whom, by hypothesis, are supposed to have the same numbers of total errors, —the sort of situation, in short, which our technique of partial correlations arrives at algebraically. And we have let curve A represent an individual for whom this given number of errors was made primarily at first; curve C, on the other hand, an individual for whom these errors were made more at the end; and curve B, an individual for whom the errors were more evenly distributed throughout the entire course of learning. Let us remember, further, the fact that as the trials continue there has been a relatively steady increase in mere *rate* of running. This is a fact which can be testified to by any operator. It seems to be due to mere habituation to the general maze situation, and to be relatively independent of specific knowledge of blinds or lack of blinds. The "stupid" rat who continues indefinitely to enter certain blinds, nevertheless speeds up as far as mere rate of running goes, just about as rapidly as does the "bright" rat who eliminates the blinds. But if such is the case, it means then, further, that an error (entering of a blind) made toward the latter part of the training period tends to take, as such, less time than

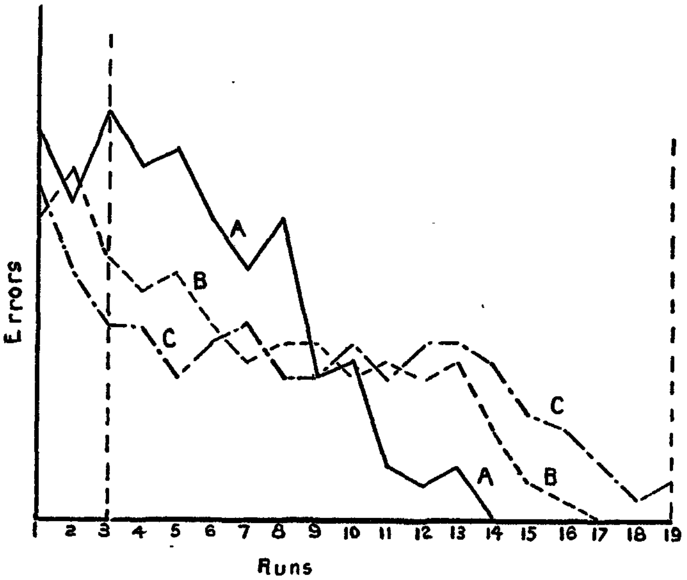


FIG. 7

an error (entering of a blind) made in the first part of the training period. Returning, then, to figure 7, this means that, although rats A, B, and C have, by hypothesis, all made the same number of errors, the errors of rat A, since they were made early in the game, probably took much more time than did, say, those of rat C which were made in larger proportion toward the end of the training period, when all rats tend, as such, to be speedier. In other words, the total time score for rat A must have been relatively large, that for rat C relatively small, and that for rat B intermediate (irrespective of which was innately more speedy, and irrespective of the fact that the total error scores of all three are by hypothesis the same). But rat A is also, as we see from inspecting the curves, the one with the largest perfect run score, and rat B the one with the intermediate perfect run score, and rat C the one with the smallest perfect run score. In other words, we now see why, when errors are held constant (as in the formula for partial correlation, or as here diagrammatically shown), there appears a *positive* correlation between time and perfect runs. It is because not all errors are the same. Errors made at the beginning of the training period have a bigger time value than those made near the end. And the positive partial correlation obtained between time and perfect runs with errors constant is but an artifact of this circumstance. This positive correlation, in short, merely points to the fact that errors take more time if made near the beginning of the training period than if made near the end.[12]

Again, we see that the time score is an ambiguous one. For it now appears that it gives more weight to errors made near the beginning of learning than to errors near the end of learning. But if *a priori* any choice is to be made, it would seem that errors in the latter part of learning should, if anything, be given the greater weight.

To sum up then: Time is an ambiguous measure for two reasons: (1) because it is a composite from two independent factors,

[12] The authors wish to acknowledge their indebtedness to Miss Barbara Burks for the point of view underlying this analysis. See her extremely important articles, already referred to (1).

(*a*) a factor of pure speediness *per se*, and (*b*) the number of errors made; and further (2) because in so far as it depends upon the second factor, errors, it gives greater weight to errors made at the beginning of the learning curve than to errors made at the end of the learning curve.

Retracings

In those mazes where doors are not used, we may have, as a fourth type of score, the number of retracings (*R*) made during the course of learning. If, now, we correlate these retracing scores with errors and time and perfect runs, we obtain the following:

$$r_{RE} = 0.687$$
$$r_{RT} = 0.923$$
$$r_{RP} = -0.359$$
$$r_{PE} = -0.838$$

To interpret these figures, let us treat retracings as we did time. Let us obtain the partial correlation retracings vs. perfect runs, with errors constant. This is $r_{RP.E} = 0.547$. In other words, retracings, like time, correlate positively with perfect runs, when errors are held constant.

To explain this result, let us note two concrete facts: (1) There is the fact that will be testified to by any careful rat operator, that retracings are made primarily in the early days of the maze experience; and (2) there is the obvious fact that the making of retracings quite simply and purely mechanically tends to raise the number of blind entrances in the particular trials in which such retracings are made. For an animal, by simply going over the path a second time (due to having retraced) has thereby another chance at again entering the blinds which he has not as yet eliminated. These two points taken together mean that the more the retracings, the more the errors will tend to be heaped up at the beginning. But this again means, when the total errors are the same, that the more the retracings, the greater the number of perfect runs. Or, in other words, we have the explanation of why retracings correlate positively with perfect runs when errors are held constant. Many retracings mean many errors at the beginning, and (when errors are held constant) many errors at the beginning mean many perfect runs.

So much for the purely theoretical consideration of why it is that high retracing scores (like high time scores) tend to mean early errors and hence (when errors are held constant) more perfect runs. The practical question which remains is what to do about it. What can we conclude from the above as to the diagnostic significance of retracings? Shall we allow retracings with their corresponding effect upon the error scores, or shall we not?

Two hypotheses as to the causes of retracing suggest themselves. The first hypothesis would be that retracings are primarily a result of nervousness and caution. But, if so, and we wish primarily to measure learning ability *per se,* it seems undesirable merely to increase the initial error scores simply because of this nervousness. The second hypothesis would be that the initial retracings, occurring at the first as they do, are rather an expression of intelligence in the sense of expressing a propensity on the part of the animal for an initial thorough examination of the field before attempting to settle down to mere routine learning. But from this point of view also it seems undesirable to penalize the animal's initial error score. In short, whether retracings be supposed to be a mere matter of temperament or one of intelligence, it seems that the resulting increase in initial error scores which results from allowing such retracings is undesirable, whenever, that is, we are interested primarily in learning *per se.* The present writers strongly recommend, therefore, whenever the interest is in obtaining reliable and valid measures of learning *per se,* the use of doors to prevent retracing. And, finally, it may be noted that this use of doors has certain purely practical advantages in that (1) it very much decreases the amount of time taken by the first few trials, and in that (2) it also makes it possible to get a rat to run through longer (and, that is, more reliable) mazes. For if doors are introduced, the animal is kept going in a forward direction and hence does not get discouraged nor tend so readily to go to pieces in the initial trials.

Perfect runs

Finally, we may briefly recall again here the discussion concerning errors vs. perfect runs as measures of ability made in the preceding section under reliability. We there decided that perfect runs (or number of trials necessary to reach a given number of perfect runs) were not as good a measure as errors because the internal consistency of this perfect run score could not be measured. We may further note, however, that in so far as we succeed in getting mazes which give high intercorrelations of odd vs. even and half vs. half for errors,—in so far, in short, as we tend to get parallel errors curves like those in figure 1,—then there will necessarily tend to be a unity (negative) correlation between errors and perfect runs. In other words, in the ideal piecemeal mazes, errors and perfect runs will measure the same thing.

Summary and conclusions

We may sum up our conclusions as regards the relative meaning or validity of the different types of scores:

1. Time is an ambiguous measure of learning for two reasons: (*a*) It depends upon temperamental and physiological as well as upon learning factors. (*b*) It is unduly weighted by initial errors.

2. Retracings are an uncertain quantity. They may be a function of mere nervousness, or they may be a function of cautious intelligence. In either case they tend unduly to weight the initial error scores. We recommend, therefore, their mechanical prevention by the use of doors.[13]

3. Perfect runs (number of trials) is a type of score which allows no internal method of testing for its reliability. It may be discarded in favor of the error score, which latter, when the maze is reliable, tends to correlate to unity with it.

4. Errors are thus left as the one desirable type of score for measuring learning *per se*.

[13] The original credit for inventing such doors belongs to Dr. H. C. Blodgett.

REFERENCES

(1) Burks, B. S.: On the inadequacy of the partial and multiple correlation techniques. Jour. Educ. Psychol., 1926, xvii, 532–540, 625–630.

(2) Carr, H.: The reliability of the maze experiment. Jour. Comp. Psychol., 1926, vi, 85–93.

(3) Heron, W. T.: The reliability of the inclined plane problem box as a measure of learning ability in the white rat. Comp. Psychol. Mon., 1922, i, no. 1, pt. 1.

(4) Heron, W. T.: Individual differences in ability versus chance in the learning of the stylus maze. Comp. Psychol. Mon., 1924, ii, no. 8.

(5) Hunter, W. S.: Correlation studies with the maze in rats and humans. Comp. Psychol. Mon., 1922, i, no. 1, pt. 2.

(6) Hunter, W. S.: Habit interference in the white rat and in human subjects. Jour. Comp. Psychol., 1922, ii, 25–59.

(7) Hunter, W. S.: Note in answer to Professor Carr's criticism. Jour. Comp. Psychol., 1926, vi, 393–398.

(8) Hunter, W. S., and Randolph, Vance: Further studies on the reliability of the maze with rats and humans. Jour. Comp. Psychol., 1924, iv, 431–442.

(9) Kelley, T. L.: Statistical method. Macmillan, New York, 1923.

(10) Liggett, J. R.: A note of the reliability of the chick's performance in two simple mazes. Ped. Sem., 1925, xxxii, 470–480.

(11) Paterson, D. G.: The Johns Hopkins circular maze studies. Psychol. Bull. 1917, xiv, 294–297.

(12) Randolph, Vance, and Hunter, W. S.: A note on the reliability of the maze as a method of learning in the angora goat. Ped. Sem., 1926, xxxiii, 3–8.

(13) Ruch, G. M.: Correlations of initial and final capacities in learning. Jour. Exp. Psychol., 1923, vi, 344–356.

(14) Ruch, G. M.: The influence of the factor of intelligence on the form of the learning curve. Psychol. Mon., 1925, xxxiv, No. 7.

(15) Tolman, E. C.: Inheritance of maze ability in rats. Jour. Comp. Psychol., 1924, iv, 1–18.

(16) Tolman, E. C., and Davis, F. C.: A note on the correlation between two mazes. Jour. Comp. Psychol., 1924, iv, 125–136.

(17) Tryon, R.: Effect of the unreliability of measurement on the difference between groups. Jour. Comp. Psychol., 1926, vi, 449–453.

(18) Warden, C. J.: Some factors determining the order of elimination of culs-de-sac in the maze. Jour. Exp. Psychol., 1923, vi, 192–210.

(19) Warden, C. J.: A comparison of different norms of mastery in animal maze learning. Jour. Comp. Psychol., 1926, vi, 159–179.