STUDIES IN INDIVIDUAL DIFFERENCES IN MAZE ABILITY[1]

I. THE MEASUREMENT OF THE RELIABILITY OF INDIVIDUAL DIFFERENCES

ROBERT CHOATE TRYON

National Research Council Fellow, University of California

The general problem upon which the writer is engaged (21) has as one of its objectives the discovery of the extent to which individual differences among rats in the learning of a T-maze are determined by genetic or hereditary factors. Before this can be satisfactorily accomplished, however, it is necessary to determine the degree to which the differences between individuals are caused by chance factors, commonly called errors of measurement. If the rôle of chance factors as differentiae is negligible, then, obviously, the differences between individuals are caused entirely by systematic factors, and it will remain for us only to determine what portion of these systematic factors are genetic.

The task of ascertaining the rôle of unsystematic chance factors in the causation of individual differences among rats will be attacked in this paper. The procedure is, in essence, that of determining the reliability coefficient of the maze measures, for from this coefficient we may discover the rôle played by errors of measurement. The problem is not only of importance to our inheritance experiment, but it should be of interest to maze ex-

[1] This article is the first of a series of studies on the topic of individual differences in maze ability among rats. A second paper will be on "The Determination of Individual Differences in Maze Ability by Age, Weight, Sex, and Pigmentation," and a third will treat of "The Community of Function between Two Maze Abilities." The experimental details will be presented in this paper, and not in the succeeding ones.

At this place the writer wishes to state his extreme indebtedness to Professor E. C. Tolman, who has supplied encouragement and aid throughout the execution of these experiments, and also to the Research Board of the University of California for defraying the costs of the experiment.

perimenters in general, for there has been considerable controversy as to whether the rat maze can be constructed so as to call forth consistent individual variation in performance.[2]

Definite measuremental principles were utilized, in constructing the mazes to be used in the inheritance problem. This paper will report upon these principles, upon how they were given experimental expression, and upon the results which were obtained by their use. The justification for presenting them at some length in the abstract and also in the specific form in which they were used in these experiments is (*a*) that the principles have not been so formulated in other writings, (*b*), that they have not been utilized in full in previous investigations, a situation which has produced conflicting results and needless controversy, (*c*) that they are *general* and may be applied to any maze (or other type of problems) where a reliable scale is desired, and (*d*) that the two experiments here reported were deliberately planned with these principles in mind, and the result was that one of the mazes, *Maze X*, rendered a reliability coefficient higher than that reported upon any human mental test of which the writer is aware, whereas the coefficient on the other maze, *Maze Y*, was quite as high as that of the best human measurements.

I. INTRODUCTION

a. Studies on the reliability of individual performance in maze ability made previous to the inception of these experiments

The Kansas studies. The work of Professor W. S. Hunter and his students has uniformly presented negative evidence as to the reliability of the maze. In 1922, Professor Hunter (7) reported an experiment in which he cited as the highest correlation, an *r* of .36 between the total net-time rats required to make and then to break a circular maze habit. Twenty-five thirty-day-old albino rats were used. Hunter concluded from this

[2] It should be noted that T. L. Kelley (11, p. 210) has set up a high, though quite arbitrary, criterion that the reliability coefficient of measures must be .96 or higher before a mental measuring stick can be deemed a suitable indicator of individual differences in the function measured.

study that the reason rats showed lower correlations than human beings is that the latter are easier to control experimentally, or rather that "the rat is more unstable than human adults" (p. 57). In the same year, Heron and Hunter (5) reported the performance of 22 fifty to seventy-day-old white rats on an inclined plane box and on a circular maze. "The highest correlation was −.33, which is between the time for the fourth trial in the problem box and gross time for the third trial in the maze" (p. 29). In another experiment in which 25, 36, and 24 thirty-day-old rats were run respectively on three mazes of increasing complexity, the highest correlation found was .69 between the last two-tenths of learning (measured by time) on the most complex maze. Most of the coefficients were much lower. Hunter concludes, again, that "the rat even in the simplest habit shows a greater instability of response than man" (p. 56). Hunter and Randolph (9) reported in 1924 some experiments in which rats were run on a problem box, a straight-away maze, and a T-maze, all of simple construction. The highest reliability coefficient (computed by the odd vs. even element method) was .69 for 19 thirty-day to forty-day-old rats on the T-maze. The correlation of straight-away with T-maze was −.16.

In summary, it may be said that the Kansas experimenters have constructed mazes which were wholly inadequate to measure individual differences in maze ability. These results they believed to be due, in the first place, to the instability of the rat (Hunter) and, in the second, to the inferiority of the maze itself as a measuring instrument. Regarding this second point, Heron (4) says, relative to his maze experiments on human beings, "In no maze can it be said that individual ability plays a dominant part" (p. 14). Because of its inadequacy as a measure of individual differences, Professor Hunter even aspersed the maze as a device with which to measure differences between groups (9, pp. 440–442). For this contention he was criticized by Carr (2).

The California studies. The experiments reported in this paper are a continuation of Professor E. C. Tolman's early work on the inheritance problem. He began his investigation in 1922, and reported his preliminary results in 1924 (15). In this report he described the performance of 82 unselected ninety-day-old rats on a maze which gave as its best reliability coefficient of total errors the value, .509, a discouraging inadequacy. To secure evidence of the validity of this maze, Davis and Tolman (3) ran 13 sixty-day-old rats on this and another maze, and the best correlation between errors made on these two mazes was .656. The number of rats were so small in this case as to make any interpretation of this correlation of doubtful value.

Surmising that the *shape* of the choice points in the maze was an important factor affecting reliability of performance, Professor Tolman then constructed T-units which could be concatenated in a multiplicity of ways.[3] Such units constituted uniform scoring points where an animal was required at every blind to make a deliberate choice between cul-de-sac and true path. In 1926, the writer became engaged with the problem, and it was on the foundation of Professor Tolman's earlier work with mazes that the later reliable mazes were built.

b. The eight cardinal experimental-statistical principles of measurement which were utilized in the execution of these experiments

The experimental evidence up to 1926 indicated, as we have seen, that apparently no maze had either reliability or validity. To get some clue as to the reason for this depressing situation, the writer examined some 140 or more basic references on the reliability of *human* mental tests. The net result of this review was to make very plain the fact that principles which had been applied in the construction of reliable human tests had not been systematically applied in maze experiments. In the physical construction, therefore, of Maze X and Maze Y, in the choice of an unselected sample of animals, in the development and execution of the experimental procedure, and in the analysis of the results, the writer used these principles, which were derived, as we shall see, mainly from the literature on human mental measurement. These principles are given below, first in their abstract

[3] In their first form, these T-units were described by Tolman and Jeffress (16).

form, then as they were neglected or misapplied in general in previous maze experiments, and finally as they were utilized here.

1. The greater the amount of "material" in a test, the greater its reliability, other things being equal.

Sixteen years before, C. Spearman (12) and W. Brown (1) had independently derived formulae demonstrating this principle. Yet with rat mazes, reliability coefficients had been interpreted without much thought as to what they would have been had the mazes been more difficult (i.e., possessed more cul-de-sacs, or had the blinds been more complexly concatenated). The T-maze material devised by Professor Tolman lent itself especially well to the application of this principle. In building Maze X, therefore, the writer placed as many of these units (17 in total) as could be put in the available maze room. After running a few trial rats, he even changed the pattern to what appeared to be one more difficult to the rats. Maze Y, made up of a similar kind of T-unit material, was made to consist of as many units (20 in all) as could be complexly concatenated in another room. Thus an effort was made to increase the material in the mazes to as great an amount as seemed physically possible in the available space.

2. The individual performance is more reliable in proportion as the individual is "test-broken," other things being equal.

By being "test-broken" one means being so adjusted to the testing situation that no variable emotional responses affect the score. In previous experiments, animals were ordinarily introduced "green" into the maze. Such a situation almost always produces a differential emotional "up-set," thus introducing a source of unreliability. Some rats, stupid in the first trials apparently because they are afraid of the apparatus, often become bright after they have lost their apparent fear. Such a switch in performance reduces the reliability coefficient of total errors.

To remove this disturbing influence, preliminary runs on a practice path, which was the same length for all animals, were given to all rats before they ran upon Maze X or Maze Y proper. This path served to acquaint the animals with the mechanical features of the maze upon which they were to run later. As a consequence, when they encountered the maze proper for the first time, they were negligibly affected in an emotional way by the mechanical features of it.

3. The more controlled the experimental situation, the higher the reliability coefficient of the scores, other things being equal.

The situation is considered more controlled in proportion as the experimental conditions are held constant for all animals. Previous maze experiments were probably as satisfactory in this connection as experiments with human tests. In one regard, however, the writer felt that an improvement may have been made, namely, in the handling of the animals before running. The customary method of pursuing the rat in his cage with the hand, grabbing him, lifting him out, and dropping him into the starting box seemed likely to occasion an unreliability of performance, especially in the earlier trials when the rat is as yet not negatively adapted to the experimental features.

A system was devised in operating Maze X and Maze Y whereby a rat was not handled for several minutes before running the mazes. With Maze X, this was accomplished by means of an automatic revolving delivery table in which the rat lived during the entire experimental period. From his own special compartment in this table, each rat entered the maze by his own volition.[4] With Maze Y, this control was accomplished by means of a "starting nest" consisting of a box of compartments, one compartment for each rat. Each section which contained a rat had one exit into the maze which was blocked until the rat was ready to be run. The rat was then liberated into the maze when the whole nest was moved in such a way that his exit came opposite a hole in the block through which he entered the maze. Thus he was not handled bodily.

4. The less subjective the scoring of the performance, the higher the reliability coefficient, other things being equal.

The truth of this principle is too well known to require elucidation. Recording the performance of a rat in the maze has gener-

[4] For a complete description of Maze X as well as of the procedure employed in its use, see the recent paper by Tolman, Tryon, and Jeffress (17).

ally been accomplished by marking on paper the errors made, or the distance travelled, or by noting the time on a stop watch. Subjective judgments are required at those moments when the observer must estimate whether the rat's random movements at choice points are errors, or whether entrances are partial, full, etc. If the distance covered is recorded, judgment is required in estimating these distances. Unreliability in starting or stopping the stop watch may likewise influence the time scores.

An almost complete objectivity was accomplished on Maze X, where the rat wrote his record on a moving tape (17, p. 107). The only possible source of error, it seems, was in failure of the apparatus to record by virtue of some break-down (a rare occurrence), and in the transcribing of the tape record onto the tabulating sheet. The performance on Maze Y was recorded by the less reliable eye-hand method. The slightly smaller reliability coefficient which was found on Y may have been due in part to this difference between X and Y in scoring objectivity.

5. The greater the "spread of talent" in the group of subjects measured, the higher the reliability coefficient, other things being equal.

T. L. Kelley (10) had in 1921 derived a formula showing the systematic increase of reliability with increased spread of talent. In maze studies, however, the intrinsic variability of the animals from a genetic point of view had been rarely considered. Whether they were inbred, litter-mates, or random selections from many unrelated individuals was infrequently reported, perhaps not even known.

To avoid homogeneity in selection, an especial effort was made to secure for these experiments a random unselected sample of animals. The rats were therefore drawn from at least 30 different litters, parents of which were not closely related. The writer was informed that no system of selective inbreeding had been used in the colonies from which this unselected sample was derived.

6. The greater the heterogeneity of correlated "irrelevant factors", the higher the reliability coefficient, other things being equal.

By "irrelevant factors," one means those in which one is not essentially interested as causes of individual differences in the mental function, and hence as a source of variation one would wish to control. As is well known, for instance, intelligence tests show higher reliabilities if one permits a flagrant heterogeneity of age (below a certain maximum), health, special training, etc. One must show either that such factors are not causal differentiae (and hence correlate zero with performance), or if they are such, reliability coefficients must be quoted for constant values of these irrelevant variables. Since rat maze measures showed such inferior reliabilities, it may seem academic to mention this type of causation of reliability. Nevertheless, no reliability coefficient should be quoted without qualifying it by a consideration of these factors, for however low the coefficients may be, the magnitudes which do arise may emanate solely from such irrelevant sources.

Consider first, age. To permit the ages of the animals to vary below about 75 days, i.e., sexual maturity, might *produce* a correlation. Just such an effect occurs in human mental tests, due to the rapid rise in the mental growth curves below 16 years. On the other hand, it may have a disturbing reducing effect on the reliability coefficient. The frequently used method of selecting rats 30 days old at the beginning of the experiment, and running them, say, for 20 or 30 days, is in effect the selection of just-weaned rats and the running of them during the period of their most rapid physical (and probably mental) growth. Different rates of growth among the rats may produce differences in performance—differences which might not have appeared had the rats been run after maturity. Thus it seems somewhat precarious to use premature rats before these matters have been investigated. Since many of the previous experiments used 30-day-old rats, here may have been a source of unreliability. The safer procedure is to use adult rats and to let their ages vary through a defined region of age, then calculate the correlation between this age variation and performance. If a significant correlation appears, then one must quote reliabilities and other correlation coefficients for the restricted ranges of age variation in which the age-performance correlation is zero. This last procedure has been used in our X-Y experiments.

Consider next, weight. Food being the customary reward, in maze-running, rats of different body weight may be differentially affected by the amounts of food received at the end of the maze. If the food is adjusted to the average rat, the heaviest rats will be under-fed, and the lightest, over-fed. Thus weight may be a "correlation-producing" factor through the medium of the reward in running. To determine if it plays such a rôle one must calculate the weight-performance correlation coefficient. If a significant correlation appears, then it must be removed, if possible, by feeding in proportion to weight. But if such a correlation cannot be removed by this experimental means, the reliability coefficient must be quoted separately for those regions of weight variation in which the weight-performance correlation is zero. In the correlation studies prior to 1926, no cognizance of this weight factor seems to have been taken. In the maze study here reported, careful weights were taken of the animals and the weight-score correlation was considered.[5]

Another factor is sex. If the mean performance of males and females is significantly different, then the reliability coefficient, even though it were zero when computed in each sex group separately, would become greater than zero when these two groups

[5] In later studies, Professor C. P. Stone has taken cognizance of the weight factor but he has used a method of controlling weight in a number of studies, e.g., The age factor in animal learning (J. Genet. Psych., v, No. 1, vi, No. 2), which does not necessarily remove, in my opinion, the influence of weight on performance. His procedure is to hold rats at a constant weight during the experimental period by daily weighing and adjusting the food for rats. In the case of growing rats, these adjustments are made, however, to permit a certain constant rate of increase in weight. Besides being very laborious and requiring additional handling of the animals, such a method does not preclude the possibility of weight factors affecting score, because it does not control the effects of variation in *rates* of growth between different individuals. For instance, two animals of the same age and weight at the beginning of the experiment may possess genetic factors for different rates of growth such that under normal feeding conditions, one rat would *eventually* be much heavier than the other. Hence, keeping them at the same weight (or permitting the same increase of weight) would amount progressively to underfeeding one animal or overfeeding the other. If the general sample of rats under investigation is very heterogeneous genetically, and begins running at an age of greatest growth (30 days) such a weight control might introduce a serious source of variation in maze performance due to genetically determined differential rates of growth.

are considered as one. This fact has been succinctly proved by Yule (23, p. 218). Maze experimenters ordinarily controlled this by using only one sex, unfortunately limiting their conclusions only to the one sex. In our X-Y experiments, both sexes were used, and the effect of sex differences on maze score determined. Another dichotomous irrelevant factor to which the same considerations apply is coat and eye pigmentation. In our maze experiments albino and pigmented animals were both used, and the influence of this pigmentation difference on performance ascertained.

7. The reliability coefficient is increased, the more "comparable" the two sets of measurements correlated, other things being equal.

The true reliability coefficient of any set of scores is defined (22) as the correlation of the set of scores with another comparable set. Strict comparability exists only when the experimentally measured material in one set is a random sample of measurements for each individual of the same general type of material as that from which the other set is drawn. When the conditions of comparability are satisfied, the *standard deviations of each set of measurements are equal.* For such experimentally determined material as scores on successive trials on a maze, Spearman (12) devised the method of collecting odd elements in one set, and even elements in the other set (the elements may be trials, or blinds), correlating these two independent sets, and then applying the correction which eventually became known as "Brown's formula". The corrected correlation is the reliability coefficient of total scores on all trials or blinds. This method, called the "split-test" method, is the only exact method of determining the reliability coefficient of maze scores. All other methods, such as that of correlating the first half of the trials or blinds against the last half, or correlating one maze against another, do not satisfy the *definition* of the reliability coefficient, for the two sets of measurements used in the other methods are rarely comparable according to the criterion of comparability. The standard deviations of such sets are, of course, rarely equal. As a consequence of lack of comparability, correlations between such sets of measures give values *lower* than the true reliability coefficients. Besides the matter of equality of the standard deviations, there are other conditions which should be satisfied if the true reliability coefficient is to be secured. By using the "split-test" method, however, one can generally be assured that such other conditions are satisfied. In calculating the reliability coefficients of our two mazes, we have used only the exact "split-test" method.

8. Finally, the fewer the subjects, the less dependable the computed reliability coefficient, other things being equal.

In theory everyone is familiar with this axiom; in practice reliability coefficients and elaborate correlation analyses have been made on samples of considerably less than 30 animals. Most of these coefficients are uninterpretable, since the conventional P.E._r is inapplicable to them. More than any other type of research, studies on individual differences involving correlation require the measurement of large samples before any sort of generalization may be drawn. In our experiments, therefore, in order that the results would be considered dependable, one hundred and forty-one animals were run on the mazes.

Except for the last, each principle described above pertains to some factor which may possess varying degrees, and which has a systematic influence on the value of the reliability coefficient. Thus, the reliability coefficient of maze measures is *higher* in proportion as (*a*) the maze has more blind alleys, or more trials are included in the maze score, or the blinds are more complexly concatenated, (*b*) the animals are more "test-broken," (*c*) the experimental procedure is more controlled, (*d*) the scoring is less subjective, (*e*) the rats are less inbred and unselectively sampled, (*f*) there exists more heterogeneity of "irrelevant" factors, and (*g*) the two sets of measures from which the reliability coefficient is computed are more comparable. The effect of the factor embodied in the eighth principle is unsystematic, for as one increases the number of subjects, the value of the reliability coefficient may go up or down, approaching eventually some "true" value as the number of cases becomes large.

These eight cardinal principles were brought over almost bodily from the principles of human mental test construction, and no great metamorphosis was necessary in the transfer. A few maze experimenters utilized some of them, but neglected others equally as important. The writer endeavored to apply all of them. As to the fruitfulness of the application of them, evidence will be cited in section III of this paper.

c. Studies in the reliability of maze performance since 1926[6]

In 1926, Professor Hunter published a reply (8) to Carr's remarks (2) on the reliability of mazes. In this reply he reiterated his previous position as to the undependability of group differences as ascertained on the maze when individual differences are unreliable. Hunter quoted from a letter from Kelley, who stated that "if the proper formula for the standard deviation of a difference is used . . . . it is unnecessary to know what is the reliability of your measure in order to know how much confidence you can place in an obtained measure of difference" (p. 397). Following Kelley's thought, the writer published a rather unsatisfactory statement (18) to the effect that Hunter's contention was contrary to statistical theory. This statement was later (19) amplified, but only after a criticism by Huffaker (6) was I able to place it in a final concise form (20). Applied to maze measures, my notion was, briefly, that when a difference between mean maze scores of two groups who differ in some systematic way *has been found* to be statistically reliable by gauging it in terms of the orthodox $\text{P.E.}_{\text{diff.}}$ formula, the reliability of this difference is unaffected by knowledge of the reliability coefficient within each group. The question as to whether a maze, completely unreliable *for the whole range* of rat talent, *will show* group differences, is a matter irrelevant to the statistical issues involved in the above controversy, which dealt with group differences which had actually *been shown* to exist.

In December, 1927, Tolman and Nyswander (24) published

[6] I have placed the review of these studies at this point for the reason that they were published after the inception of this experiment. I wish to state, however, that although the publications of Tolman, Nyswander, and Stone came later, these authors materially contributed from their knowledge and experience with mazes to the X-Y experiments here reported. From Professor Tolman, especially, under whom the writer worked as a graduate student, constant helpful advice and criticism was received.

an analysis of the reliabilities of a number of mazes of different shapes. These writers realized clearly and specifically mentioned principles 1, 5, and 3 described above. They reported a multiple-T maze which gave a reliability coefficient of .909 on a sample of 25 rats. Another experiment, described by Stone and Nyswander (13) embodied principles 2, 5, 6, and 8, and hence their 14 blind multiple-T maze gave a reliability coefficient of .958 for errors. Several of their methods of calculating the reliability coefficient (based on the correlation of one-half the maze against the other half, and one-half the trials against the other half) probably violated the definition of the reliability coefficient. Stone later (14) analysed a modified Carr maze and deemed it as less reliable than the multiple T-maze.

II. EXPERIMENTAL DETAILS

These particulars will be described under the heads of the first six principles mentioned above, but only features not stated above will be given below.

1. Difficulty (as indicated by pattern and length) of the mazes

For diagrams showing the maze pattern of X and Y see figures 1 and 2. A complete description of the details of the construction of the automatic Maze X is given elsewhere (17). The T-units of Maze Y differed from those of X in that the floors of Y were solid, while those of X balanced upon a central fulcrum and dipped slightly forward as the rats ran over them.

In the Y T-units, the stems of the T's were 22 inches long, the blind and true path arms of the T were together 36 inches. The alleys were 3½ inches wide, and 5 inches deep. Half way in the blind (as well as in the opposing true path) hung simple black cloth curtains which prevented the rat seeing the end of the blind at the choice point. As the rat entered the stem of each T-unit he walked upon a wire door, which was hinged at the bottom and extended up and away from him at an angle of about 45°. The door was held up by a rubber band in such a way that as the rat walked upon the door it gave under his weight. As he walked

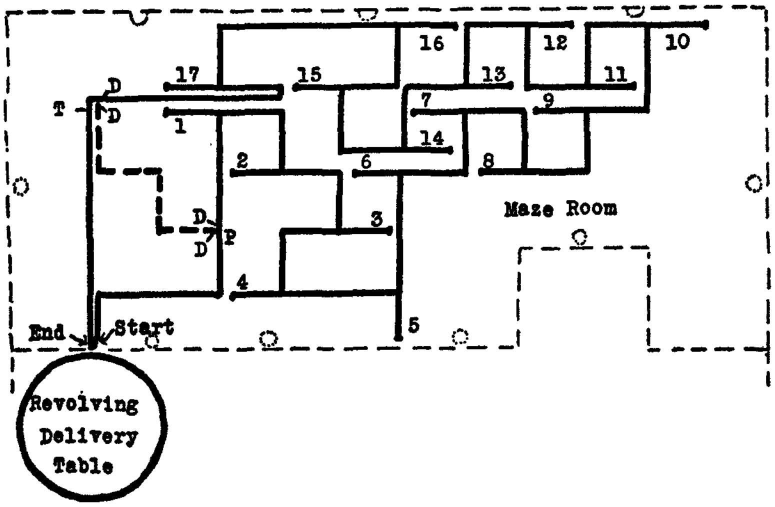

FIG. 1. DIAGRAM SHOWING PLAN OF MAZE X

Numerals are blind alleys; dotted circles are overhead indirect lights; and D's are hand-operated doors.

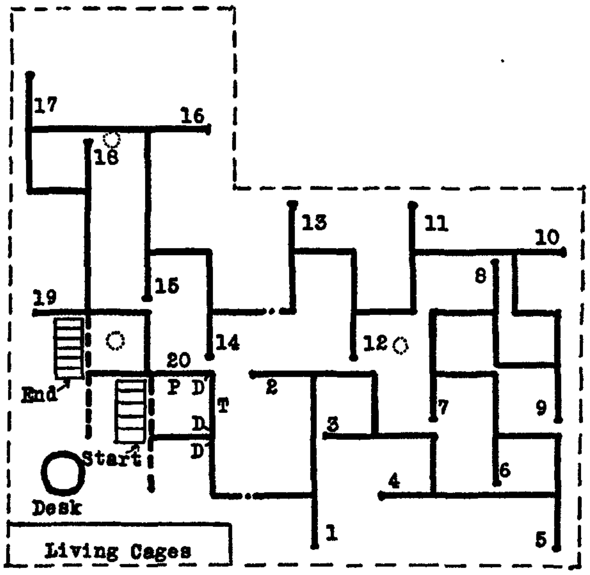

FIG. 2. DIAGRAM SHOWING PLAN OF MAZE Y

Numerals are blind alleys; dotted circles are overhead indirect lights; and D's are hand-operated doors.

over the door, it lay flat beneath him and then was pulled up behind him by the elastic band as soon as he stepped off. Thus he could not retrace back into the last unit which he just left.

2. The "test-breaking" practice

The practice path upon which the rats ran before encountering the Maze X proper is shown in figure 1. It begins at "Start," then goes to *P*, *T* and "End." The procedure of running is described elsewhere (17). Briefly summarized, it may be said that the animal had short trips over the practice path for eight days before running X proper, and he was introduced gradually to curtains and doors in this path. Regarding the practice path preliminary to Maze Y, this goes (see figure 2) from "Start," through *T*, *P* and to the end. The rat was first left without food over-night in his starting compartment in order that he may be more accustomed to his new surroundings. On the next day, he made three runs through the preliminary path, and on the following day two trips through it, these followed by his first run on Maze Y proper.

3. Control of experimental conditions

a. Of the environs. Maze X was lighted indirectly by eight 60 watt lights, Y by three 200 watt globes. Both maze rooms were isolated in the basement of the building. No one was present in the maze room when the rats' ran on Maze X, but during the Maze Y runs, the experimenter sat at the "Desk." Since the rooms were in the basement, the temperature was uniformly cool.

b. Of the procedure of running. Each rat ran one trip a day for twenty days (on X during the morning, on Y during the afternoon, but, of course, for the same rats not concurrently). The animals were not handled immediately before or after running either maze. In the Y experiment, the animals resided in the "Living Cages" (see fig. 2). When the time came for the day's run, the rat was transferred by hand to his compartment in the nest of compartments marked "Start." This nest contained compartments for five animals, the partition between them being made of wire. When a lid was put over the nest, the only point of egress was at the front toward the maze. At this side, the nest butted against a solid partition (denoted by a dashed line in fig. 2), which had one hole cut in it through which the animals could escape into the entrance of the maze. When the rat was ready to start, the nest of compartments was pulled by a cord in such a way that the compartment came opposite the opening in the partition, and through it the animal went into the maze. At the end of the maze, the rat went from unit 20 into an "End" compartment in a nest similar to that at the "Start." After the rat entered, the experimenter pulled a string in such a way that a solid partition now blocked the entrance into the "End" nest, thus preventing the rat's escaping back into the maze again. This arrangement of "Start" and "End" nests had the special advantage of not requiring the experimenter to get out of his seat during the running of at least five animals. On Maze X, an arrangement analagous to that above for Y was automatically executed (17).

c. Of the incentive. The reward was a modified Steenbock ration composed as follows: ground corn or wheat (76 per cent), linseed oil (16 per cent), casein (5 per cent), alfalfa (2 per cent), salt (0.5 per cent), calcium (0.5 per cent). This ration was made rather wet for the X maze (1½ parts of water were added), for it contained all of the water which the animals were to receive during the experimental period. All the males were fed the same amount, about 40 grams, the females were given about 30 grams. Fresh lettuce was given once a week. A different control was adopted in the Y situation: animals were permitted to eat at a filled food pan in their end chamber until they definitely turned away from the food pan.[7]

[7] The general execution of the experiment was controlled in the sense that the writer alone ran all of the animals on both mazes during the entire experimental period. The season was not controlled, for the first animals began running March 5, 1927, and the last finished August 17, 1927.

4. Scoring of the performance

On Maze X the animals registered their path through the medium of an electrical device, which recorded partial and full entrances and retracings within a unit, and time (17). Seated at the "Desk" in the Maze Y room, the experimenter recorded the rat's path by hand. In the treatment of results, partial and full entrances are not distinguished, each being called an error. No account was made of retracings.

5. The sample of animals used

The rats came from three sources: (*a*) the "8-1-26" group of 46 albinos from 9 litters (the 18 parents of which came from the Psychology colony and were largely unrelated for several generations previously), (*b*) the "9-15-26" group of about 40 unselected vari-colored males (hoods, blacks, and greys) from the Anatomy colony, and (*c*) about 50 offspring from 8 different females from (*a*) and (*b*) bred to 5 different males. In the stocks from which these animals had been derived, wild blood had on several occasions in previous years been introduced, and there was no record of inbreeding in the perpetuation of the stocks. Thus this sample was considered to be rather heterogeneous and unselected.

6. Heterogeneity of irrelevant factors

The animals were distributed in various catagories as follows:

	ALBINO	PIGMENTED	TOTAL
Males	48	40	88
Females	49	4	53
Total	97	44	141

a. Sex. The table states the numbers in each sex.

b. Pigmentation. Rats that showed any degree of hoodedness were termed pigmented, and all such animals had pigmented eyes. The albinos were white with pink eyes.

c. Age. The ages of the animals were calculated as the number of days from birth to the first day of running the maze. These ages varied as follows:

MAZE	AGE (IN DAYS)																TOTAL
	114	129	144	159	174	189	204	219	234	249	264	279	294	309	324	339	
X	17	0	9	5	11	13	5	4	4	11	12	11	8	11	6	13	140

MAZE	AGE (IN DAYS)																	TOTAL
	149	164	179	194	209	224	239	254	269	284	299	314	329	344	359	374	390	
Y	8	12	6	5	11	5	7	4	6	11	11	20	12	9	4	8	1	140

All of the ages were correct within a possible error of one or two days, except the vari-colored "9-15-26" males whose ages were not in error more than one age category. In the frequency tables here (and below), the numbers in the first row are not mid-points but denote the upper limit of the category. The numbers in the second row represent the frequencies in the categories.

d. Weight. Though weights were taken on several occasions the only ones considered here are (*a*) weight Z_1 taken before the last practice run on the preliminary path of Maze X, and (*b*) weight Z_2 taken before the last (20th) run on Maze Y. These are the absolute weights, therefore, at the beginning and at the end of learning. In order to analyze the effect of weight *alone* on score, it was necessary to hold sex constant. I have therefore presented the weight data for males only, the females not having a sufficiently large number of cases to justify analysis. These weights varied as follows:

WEIGHT	WEIGHT (IN GRAMS) OF MALES ONLY																TOTAL
	184	199	214	229	244	259	274	289	304	319	334	349	364	379	394	409	
Z_1	1	1	2	3	9	6	14	16	13	7	8	5	2	0	0	1	88

WEIGHT	WEIGHT (IN GRAMS) OF MALES ONLY														TOTAL
	189	199	209	219	229	239	249	259	269	279	289	299	309	319	
Z_2	3	4	4	3	8	11	10	17	8	5	6	6	2	1	88

e. Interval which lapsed between running X and Y. The regular procedure called for a rest of about five days after running X and before beginning Y. During this time, the animals had food and water continuously before them. Not counting the preliminary practice before Y, an interlude of 8 days (with a possible error of one day) elapsed between the last X run and the first Y run. One hundred and ten animals, termed the "Regular" group, experienced this interval. Certain delays, however, resulted in an "Irregular" group of 31 animals, which experienced interludes as follows: 23 days (9 rats), 28 days (5 rats), 36 days (7 rats) and 56 days (10 rats).

III. THE DETERMINATION OF THE RELIABILITY OF INDIVIDUAL DIFFERENCES ON EACH OF THE MAZES

Our results consist in the calculation of the reliability of individual differences in the two scores:

X = total errors made on trials (days) 2–19 on Maze X
Y = total errors made on trials (days) 2–19 on Maze Y

1. The reliability coefficients

The reliability coefficient of X is based on the correlation of the sum of errors made on even trials ($X_e = 2 + 4 + \ldots. + 18$), against the sum of the errors made on odd trials ($X_o = 3 + 5 + \ldots. + 19$). The assumption made is that these sub-variates are comparable measures of X, i.e., are similar samples of the learning measured by X. On logical grounds it seems evident that X_o and X_e are such samples since they represent random selections through the X data.[8] The same reasoning applies to the Y data, where the same method of sectioning was used.

[8] Random sectioning is not an infallible procedure of securing comparable samples. Mr. Sidney Adams, working with the writer's data, has calculated the reliability coefficients based on the correlation of errors on odd vs. even *blinds*. His coefficients lie below the coefficients based on trials for the reason probably that the odd blinds are not strictly comparable to the even. The total number of blinds are too few to insure that, on X, for instance, the odd 9 are strictly similar to the 8 even. By chance, one set may be more difficult and measure somewhat different behaviors than the other. Thus the odd vs. even *blind* correlation would be lower than odd vs. even *trial* correlations since the sub-variates would not be so comparable.

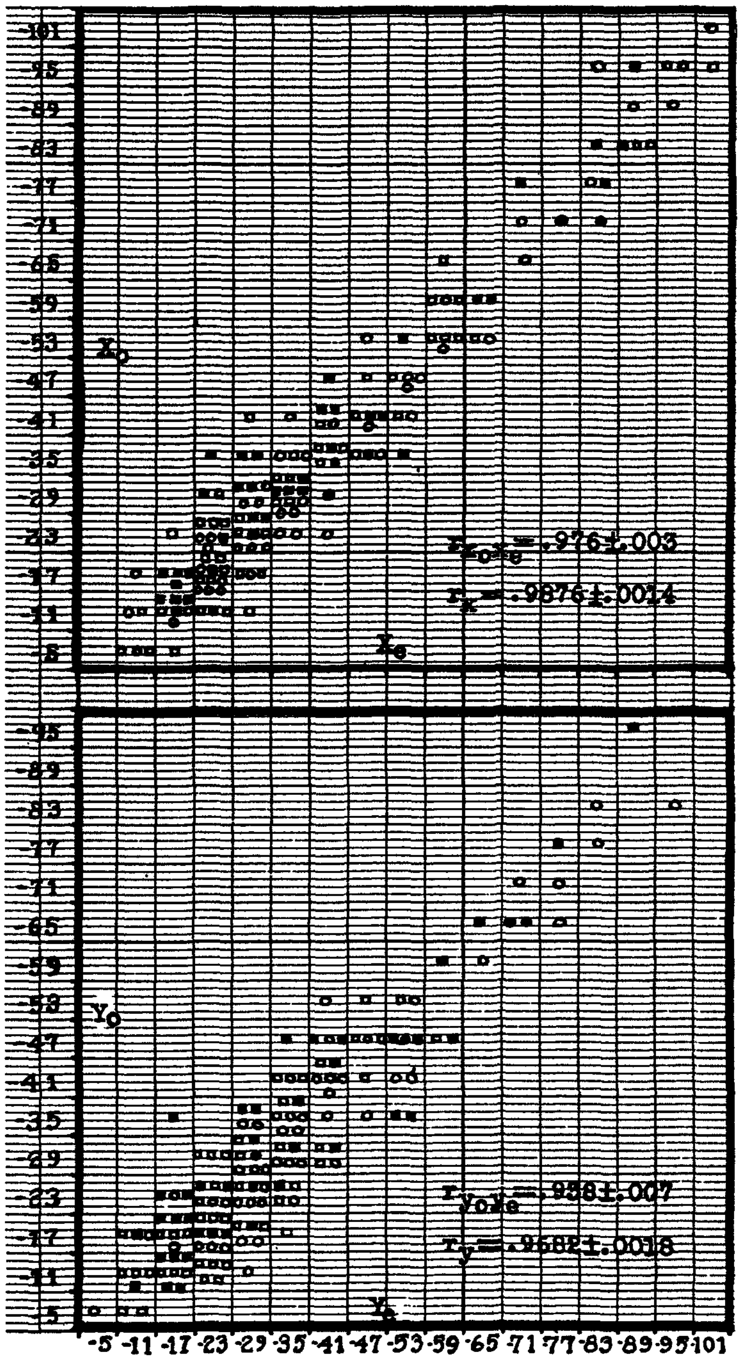

FIG. 3. SCATTER DIAGRAMS SHOWING CORRELATIONS BETWEEN SCORE ON ODD TRIALS (ORDINATES) AND SCORE ON EVEN TRIALS (ABSCISSAE)

Upper plot refers to odd-even correlation on Maze X, lower refers to that on Maze Y; scale values on the axes refer to errors (entrances into blind alleys); the abscissae at the bottom refer to both plots; entries are males (squares), females (circles), pigmented rats (blacked), and albinos (non-blacked).

When the two sub-variates upon which the reliability coefficient is based are comparable, then their standard deviations are equal, within their sampling errors. That such conditions hold with the sub-variates used in calculating our reliability coefficients is indicated below:

$$\text{S.D.}_{x_e} = 22.6 \pm .9 \qquad \text{S.D.}_{y_e} = 17.7 \pm .7$$
$$\text{S.D.}_{x_o} = 22.8 \pm .9 \qquad \text{S.D.}_{y_o} = 17.8 \pm .7$$

To illustrate the high magnitude of the correlation between odd vs. even scores, and to show that it extends throughout the whole range of talent and is not a function of a few extremely bright or dull rats, the correlation plots are given in figure 3 for X and Y respectively. The entries in these plots have been differentiated as follows: males (squares), females (circles), pigmented (blacked), and albinos (not blacked). The purpose of making these distinctions is to indicate that *the correlation is high for each of these sub-samples separately* and is a negligible function of the differences between them. The correlation coefficients were:

$$r_{x_o x_e} = .976 \pm .003$$
$$r_{y_o y_e} = .938 \pm .007$$

Since these coefficients[9] represent the reliability coefficients of only half the maze measures, it is necessary to substitute them into "Brown's formula" in order to discover the coefficient of the total maze scores. Such substitution gave the following values:

Reliability coefficient of X: $r_x = .9876 \pm .0014$
Reliability coefficient of Y: $r_y = .9682 \pm .0018$*

* Shen's P.E.'s.

[9] All correlation coefficients are Pearsonian, and all figures after ± are P.E.'s. The value, $r_{x_o x_e}$, was computed three ways, namely, from two independent scatter diagrams, both of which gave the value .975, and from the crude scores, the last method giving the value above. The value, $r_{y_o y_e}$, was computed from two independent scatter diagrams, and gave both ways the value above.

Whatever discrepancies that may appear between these values above and other values tabled in the papers which follow this are due to the fact that all values were originally calculated to the fourth decimal, but in preparing them for publication, only that place was kept indicated by one-half the probable error.

That the reliability coefficient of Maze X is higher than that of any human intelligence test appears from the recent test summary by Kelley (11, see Chap. 10: "Classified and graded lists of tests, giving reliability and other information"). From this list I have copied the highest reliability coefficients of the best intelligence tests as obtained on apparently unselected samples of children where there has been no flagrant spread due to heterogeneity of irrelevant factors, that is, as obtained upon children *of the same age* (within a year, except adults) and for a *single grade range.* These were: .93, .97, .93, .80, .85, .60, .90, .89, .91, .85, .96. The reliability coefficient of X is better than any one of the above values, that of Y is equal to the best one of them.

Several objections may be offered against comparing the rat reliability coefficients with the human. The first is that the traits compared are not analogous and cannot be set off against each other. The intention here is not, however, to prove that with reference to the *same ability* to learn in rats and in human beings the rats are more reliable. It is rather to show that, contrary to previous opinion on the matter, the rat is highly stable in performance when the stage is properly set for reliable measurement, and that, in fact, individual differences in a particular rat ability are just as reliable and systematically determined as are individual differences in certain human traits, and perhaps are more so. Whether the behaviors are identical is irrelevant. The second objection which may be offered is that the difference in experimental time alone accounts for the difference between the coefficients. With human beings, it is argued, mental tests require only about *an hour* to administer, while with rats the experimental period is much longer, in fact, 18 *days.* The answer to this objection is that the significant cause of high reliability cannot be the difference in time for the reason that previous rat experiments showed very low reliabilities for periods of measurement equally as long as that in our experiments. Furthermore, the actual time during which the rat was scored was on 18 *trials.* Now, the average trial took at most about 3 minutes, therefore, the average time on 18 trials was not longer than *one hour,* an amount of time quite similar to that allotted in human testing. The third objection would be that the rigorous age (or grade) selection imposed on the human groups caused the reliability of the human measures to be lower than that of the rats among whom a wide spread in age was allowed. But the reason such restriction has been put on human samples is that it has been found that age correlates highly with ability, so to remove the spurious effect of such maturation, reliability coefficients are quoted not for a constant age (for it is an impossibility to hold age from birth to testing constant to the very second) but for that *range* in age, generally about a year, within which the correlation between age and ability is approximately zero. Exactly the same principle has been observed in quoting our rat reliability coefficients, for it has been found that for the age range permitted in our experiments there exists a correlation between age and ability not significantly different from zero. The evidence regarding this correlation as well as the correlation between ability and each of the other irrelevant factors mentioned earlier in this paper will be given in a following paper. Suffice it is to say here that none of these irrelevant factors occasions the high reliability coefficients given above.

2. Variation of an individual rat's score in X and in Y due to errors of measurement

Having now ascertained the reliability cofficients, the next step is to determine how much the average rat would probably fluctuate in his score as a result of unsystematic errors of measurement. What we wish to know, in other words, is the magnitude of difference between two individual rats' scores which might arise by chance, and which we cannot definitely assign to underlying systematic causal factors. A measure of this "error of an individual score" is the familiar

$$\text{P.E}_X = .6745\ \text{S.D.}_x \sqrt{1 - r_x} \qquad \text{P.E}_Y = .6745\ \text{S.D.}_y \sqrt{1 - r_y}^{*}$$

* For a very lucid account of this concept, see Kelley (11, pp. 146–181).

The standard deviations of the total *X* and *Y* scores are:

$$\text{S.D.}_x = 45.1 \pm 1.81 \qquad \text{S.D.}_y = 34.6 \pm 1.39$$

By substitution of these sigmas and the proper reliability coefficients into the above probable error formulae, the probable error of an individual's score on *X* and *Y* are respectively:

$$\text{P.E.}_X = 3.38 \qquad \text{P.E.}_Y = 4.15$$

The chances are even, therefore, that the average rat would get a score ±3.38 errors away from his obtained *X* score, and so small as to be negligible that he would obtain by chance a score 4 times ±3.38 or ±13.52 errors away from his obtained score. Thus rats who differ in *X* scores by a magnitude of 13.52 errors are on the average unquestionably different from each other.

The figures are of extreme value in any experiment in which one wishes to know the accuracy of placement of the individual rats. Just such an experiment is the inheritance problem mentioned in the first paragraph of this paper. Two important uses of these figures are as follows: (A) When the experimenter wishes to choose two rats as prospective mates, he knows that in general the chances are even that two rats 3.38 errors apart are different due simply to unsystematic errors, and that it is a remote possibility that two rats more than 13.52 errors different are truly equal by virtue of chance factors. (B) The probable error of an individual score leads one to the value of the *sigma* (S.D.) of a score due to unsystematic factors (since P.E. equals 0.6745 times sigma). This sigma would be the standard deviation of individuals identical in their determination by systematic factors (22). Hence, since one of the objectives of the inheritance experiment is to establish by selective breeding a pure homogeneous line of "Bright" rats, i.e., rats identical for systematic genetic factors, then this sigma due to errors of measurement tells the minimal standard deviation of a pure line which one could possibly expect. The sigma of scores in *X* due to unsystematic factors is 5.01. Selective breeding in a group whose standard deviation was 5.01 would be of no avail, for the differences between individuals in such a group would arise solely by chance errors of measurement, and not possibly by such systematic factors as genetic causes.

IV. SUMMARY AND CONCLUSIONS

1. Derived mainly from the theory of human mental measurement, eight cardinal principles which should be utilized in the construction of reliable scales of maze ability in rats have been formulated.

2. Two T-mazes, X (17 blinds) and Y (20 blinds) were constructed, animals selected, procedure planned, and results analyzed on the basis of these principles.

3. The reliability coefficients of the maze scores (total errors made on 18 trials) were as follows: $r_x = .9876 \pm .0014$, and $r_y = .9682 \pm .0018$.

4. On the basis of these coefficients, the error of an individual rat's score due to unsystematic factors (errors of measurement) was calculated, and its practical experimental application shown.

5. It is concluded, therefore, that these principles have demonstrated from a pragmatic point of view their value, and further, in the demonstration they have indicated that the *rat maze* may be so constructed as to be *even more reliable* as an instrument with which to measure individual difference in a behavior trait *than the best human mental measuring devices.*

REFERENCES

(1) Brown, W.: Some experimental results in the correlation of mental abilities. Brit. Jour. Psychol., 1910, iii, 296–322.

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

(3) Davis, F. C., and Tolman, E. C.: A note on the correlation between two mazes. Jour. Comp. Psychol., 1924, iv, 125–135.

(4) Heron, W. T.: Individual differences in ability versus chance in the learning of the stylus maze. Comp. Psychol. Monog., 1924, ii, no. 8, 60.

(5) Heron, W. T., and Hunter, W. S.: Studies of the reliability of the problem box and the maze with human and animal subjects. Comp. Psychol. Monog., 1922, i, no. 1, 56.

(6) Huffaker, C. L.: Effect of errors of measurement on the difference between groups. Jour. Comp. Psychol., 1928, viii, 313.

(7) Hunter, W. S.: Habit interference in the white rat and in human subjects. Jour. Comp. Psychol., 1922, ii, 29–59.

(8) Hunter, W. S.: A reply to Professor Carr on "The Reliability of the Maze Experiment." Jour. Comp. Psychol., 1926, vi, 393–398.

(9) Hunter, W. S., and Randolph, V.: Further studies of the maze with rats and humans. Jour. Comp. Psychol., 1924, iv, 431–445.

(10) Kelley, T. L.: Reliability of test scores. Jour. Educ. Psychol., 1921, xi, 370.

(11) Kelley, T. L.: Interpretation of Educational Measurements. 1927, World Book Co.

(12) Spearman, C.: Correlation calculated from faulty data. Brit. Jour. Psychol., 1910, iii, 271–295.

(13) Stone, C. P., and Nyswander, D. B.: The reliability of rat learning scores from the multiple T-maze as determined by four different methods. Ped. Sem. and Jour. Genet. Psychol., 1927, xxxiv, 497–524.

(14) Stone, C. P.: The reliability of rat learning scores obtained from a modified Carr maze. Ped. Sem. and Jour. Genet. Psychol., 1928, xxxv, 507–519.

(15) Tolman, E. C.: The inheritance of maze-learning in rats. Jour. Comp. Psychol., 1924, iv, 1–18.

(16) Tolman, E. C., and Jeffress, L. A.: A self-recording maze. Jour. Comp. Psychol., 1925, vi, 455–463.

(17) Tolman, E. C., Tryon, R. C., and Jeffress, L. A.: A self-recording maze with an automatic delivery table. Univ. Calif. Publ. Psychol., 1929, iv, no. 7, 99–112.

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

(19) Tryon, R. C.: Demonstration of the effect of unreliability of measurement on a difference between groups. Jour. Comp. Psychol., 1928, viii, 1–22.

(20) Tryon, R. C.: Errors of sampling and of measurement as affecting difference between means. Jour. Comp. Psychol., 1929, ix, 191–195.

(21) Tryon, R. C.: The genetics of learning ability in rats—a preliminary report. Univ. Calif. Publ. Psychol., 1929, iv, no. 5, 71–89.

(22) Tryon, R. C.: The reliability coefficient as a per cent, with application to the correlation between abilities. Psychol. Rev., 1930, ii, 140–157.

(23) Yule, G. U.: An introduction to the theory of statistics. C. Griffin and Co., Ltd., 1922.

(24) Tolman, E. C., and Nyswander, D. B.: The reliability and validity of maze-measures for rats. Jour. Comp. Psychol., 1927, vii, 425–460.