PSYCHOLOGICAL BULLETIN
Vol. 53, No. 5, 1956

# MASS SCREENING AND RELIABLE INDIVIDUAL MEASUREMENT IN THE EXPERIMENTAL BEHAVIOR GENETICS OF LOWER ORGANISMS

JERRY HIRSCH[1]
*Columbia University*
AND ROBERT C. TRYON
*University of California*

At our present stage of ignorance about how genes determine behavior, we might well concentrate on experimental studies of lower organisms. Their reactions may be thought of as the emergent behavior which has developed through evolution into the complex behaviors of higher organisms. Knowledge gained from such studies may provide conceptual models leading to an understanding of how hereditary and stimulus components interact in determining higher forms of behavior.

For this purpose the use of lower organisms offers distinct advantages. There is a brief time span between generations, permitting *E* to perform in a short time period the various crossings essential to fundamental genetic studies. Each generation produces abundant progeny, enabling *E* to recover the extreme behavior types required in selective breeding experiments. And further, the genetics of their morphology is better understood than is that of higher forms. The fruit fly, *Drosophila*, has all of these advantages.

First, however, reliable techniques for measuring individual differences (hereafter referred to as *ID*s) in behavior must be developed. Reliability coefficients *must* be calculated, and they must be *high*. The problem reduces to the question: How can we observe the behavior of large numbers of very small *S*s and at the same time reliably measure the performance of each *S*?

This paper presents a method which accomplishes both these objectives. We call it the method of "*mass screening with reliable individual measurement.*" As an illustration of the method, we will show that in the mass observation of a particular behavior of *Drosophila*, reliability coefficients of about .9 can be secured in an experimental test period of four minutes. During this time 15 sample observations of 15 sec. each were made. Each individual was observed as a member of a group of other flies. The method shows that *Drosophila ID*s can be measured as reliably as human *ID*s. Indeed, we know of no experiment on men covering 15 brief observations that yields a reliability as high as .9.

Genetics has up to the present concerned itself with physical characteristics rather than with behavior. The reliability of individual measurement is not so obviously important in the study of morphological characteristics; usually the characteristic is either present or absent, or present in only a small number of forms, and its presence or absence is immediately obvious, (e.g., eye color, notched wing, bar eyes, etc.). Individual differences in behavior, on the other hand, are not so easily recognized:

[1] National Science Foundation Postdoctoral Fellow at the University of California, Berkeley.

such recognition requires special methods.

There are at least three reasons why we need reliable measurement of such *ID*s:

1. Reliable phenotypic differentiation is needed for selective breeding for homozygous lines. Both the purity of different strains and the rapidity of selection are limited by our capacity to discriminate between individuals, since, as the errors of measurement decrease, the probability increases that individuals with the same score will be genetically similar.

2. The study of learning also requires reliable individual measurement because of the relation between the strength of the unconditioned response and conditioning.[2] (Obviously for those individuals in whom the unconditioned response has zero strength, conditioning is impossible.) We believe that the study of learning requires reliable knowledge of the distribution of *ID*s in the population being sampled. Much effort has been spent in demonstrating the influence of environment on behavior. It is patent, however, that environmental influence must be an influence on something and therefore the laws of such influence must differ as the object influenced differs.

3. Reliable individual measurement is essential for answering three questions about the generality of any behavior: (*a*) Temporal generality; how long does a given disposition to respond endure and to what extent does the rank ordering of individuals persist over this period? (*b*) Stimulus generalization; over what range of stimuli can the response be evoked and how well is the rank ordering of individuals maintained over that range? (*c*) Behavior generality; to what extent do other behaviors preserve the rank ordering of individuals?

Efficient methods of observation are also a desideratum for studying small organisms. It is a theorem in sampling theory that the detection of extreme cases, a necessity in genetic selection experiments, requires the observation of large numbers of *S*s since the probability of finding these extreme cases is a direct function of the sample size. Rapid observation permits the examination of large numbers of *S*s and thus increases the sampling stability essential to the generality of the findings. Furthermore, replication of experiments can be undertaken without excessive labor.

The next section of this paper presents a method for reliably measuring *ID*s in behavior by means of *mass screening*, a procedure which achieves the objective of reliably *classifying every individual's behavior without handling or observing each small organism individually*. The method is completely general and easily applicable to the study of any behavior, both unconditioned and conditioned.

This objective is illustrated by the results of an experiment that employed the mass screening technique in the study of the geotropic reactions of *Drosophila melanogaster*. A series of 15 successive mass screenings, for example, produced 16 test tubes, each containing a different geotropic class of *Drosophila*. The flies in the tubes 0 to 15 represent different degrees of the negative geotropism. That is, the flies are differentiated on this final composite 16-point scale based on 15 prior mass screenings in which the individuals were not separately han-

[2] Use is made of conditioned response terminology for convenience of exposition. It is not intended to represent a theoretical statement about the nature of the learning process.

dled. The reliability coefficient of this final scale score is determinable and in principle, it can be increased to any desired value by further mass screenings.

## Experimental Design and Analytic Procedures

The method consists of cumulating a total composite score $X_t$, for each organism in any behavior, $X$, where:

$$X_t = X_1 + X_2 + X_3 + \cdots + X_n.$$

$X_1$, $X_2$, $\cdots$, $X_n$ represent scores earned by it in $n$ comparable sample mass screenings. Setting up such a total score is the essence of psychological test theory. Most of the formulae used in this paper are standard in psychological test theory. A simple summary of them can be found in J. P. Guilford's *Psychometric Methods*, Chaps 13, 14, 15 (1). Guilford's rationale of the formulae, however, is based on the factorial truth-error doctrine. In another paper one of the authors develops them with fewer assumptions (3). Our procedure adapts these principles to the problem of calculating reliability coefficients for the scores of individuals who are only observed as members of a large group.

The main steps of the procedure are as follows:

1. Conceptualize the behavior property, $X$, that is to be scaled, and operationally define it with sufficient specification to indicate the general conditions under which it may be observed.

2. Devise a standard test sample procedure for obtaining a unit measure of *ID*s in $X$, one which has the advantage of permitting observation of a large group of *S*s at one time while locating the total, $N$, of individuals in subgroup classes scored 0, 1, 2, $\cdots$, $k$ in magnitudes of $X$.

3. Take a randomly bred sample of the *S*s and mass screen them through $n$ replications of the standard procedure. At the end of every replication, score each subgroup by its cumulative total score, $X_t$, then combine subgroups with the same $X_t$ score and proceed with the next replication.

4. Calculate the reliability coefficient, $r_{tt}$, of each successive $X_t$ score, decide on the value of $n$ which will yield a reliability of sufficiently high magnitude, then examine the shape of the distribution of the $X_t$ scores of the individuals.

5. If the original method results in a low reliability or an excessively skewed distribution of final composite scores, alter the standard test, take a second random sample and repeat the general procedure. Several such experiments may be required before an adequate method of observation is discovered.

The details of the steps of this *general* procedure will be developed and illustrated by an experiment conducted by one of the authors on *ID*s in the geotropic reaction of *Drosophila.*

### 1. *Conceptualization and Definition of the Behavior*

The behavior chosen was the unconditioned disposition to go in the direction opposite to gravity. This negative geotropism is operationally defined as an upward movement of the fly whenever it is placed in any situation permitting travel upward, other external stimuli which might induce vertical movement being controlled.

### 2. *Standard Test Sample Procedure*

The test situation consists of two test tubes, a lower one standing upright in a rack, the other inverted over the mouth of the lower one. Since the flies are also phototropic,

the light source was placed at right angles to the vertical. A group of flies are placed in the lower tube, shaken to the bottom, and then allowed to ascend. At the end of an arbitrary "cutting point" time of 15 sec., a card is inserted between the lower and the upper tubes. The upper tube is scored and labeled "1," and lower tube "0."

Thus the standard sample observation in this case is like a dichotomous test item, the top tube scored "pass" and the lower one "fail." A cutoff point of 15 sec. was found empirically to divide the group of flies into two approximately equal pass and fail subgroups, a division which avoids skewness in the distribution of final composite $X_t$ scores.

It should be emphasized that dichotomous scoring is *not* a necessary restriction of the method. The standard procedure could have been devised to provide more classes. The pass-fail break was chosen for experimental convenience.

This standard test procedure, though satisfying the operational definition of geotropism, might not elicit uniquely a systematic reaction to gravity. Since the test tube situation permits only movement upward it may be that, if there is an *activity* differential among the *S*s, the flies that are upwardly mobile may be very active flies. Only additional experiments which control activity can resolve the matter. Thus, we use the term "geotropism" here only in an operational sense, recognizing that the *ID*s observed in this situation might later be shown to be significantly influenced by additional components.

### 3. *Choice of an Unselected Sample*

Since the range and reliability of *ID*s is partly a function of the heterogeneity of the *S*s, a stock of unselected *Drosophila* with a history of random mating was chosen.

### 4. *Mass Screening*

A random sample of 106 flies was screened and scored by the following procedure.

*a. First composite score,* $X_{t_1}=X_1$. The results of the first observation are shown in Fig. 1, which reproduces part of the score sheet actually used. Under $X_1$ and $f_1$ it can be seen that 54 flies ascended to the upper tube, earned a "pass" and thus received a score of $X_1=1$. There are 52 flies that remained in the lower tube, earned a "fail" and received a score of $X_1=0$. The scores, $X_{t_1}$, of this trial take the values of 1 and 0.

*b. Second composite score,* $X_{t_2}=X_{t_1}+X_2$. The 54 flies with $X_{t_1}=1$ were put through the standard procedure a second time for Trial 2. The 46 flies that ascended earn a tube score, $X_2=1$, and a composite score $X_{t_2}=2$; the 8 remaining down have $X_2=0$ and $X_{t_2}=1$, as shown. In similar fashion the flies with $X_{t_1}=0$ divide into 22 earning $X_2=1$, $X_{t_2}=1$ and 30 earning $X_2=0$, $X_{t_2}=0$.

*c. Third composite score,* $X_{t_3}=X_{t_2}+X_3$. The standard procedure is repeated for each of the three $X_{t_2}$ classes resulting from Trial 2.

Note, even though there are four $X_2$ tubes of flies at the end of Trial 2, there are only three $X_{t_2}$ classes. The two subgroups with 8 and 22 flies have been combined in one tube because both received the same score, $X_{t_2}=1$, i.e., the same composite score is the cumulative sum of all previous scores irrespective of the order in which the individual "passes" and "fails" were obtained.

*d. Additional composite scores,* $X_{t_4}$, $X_{t_5}$, $\cdots$. The procedure is continued by taking further sample observations; at the end of each one, subgroups having the same $X_t$ score are

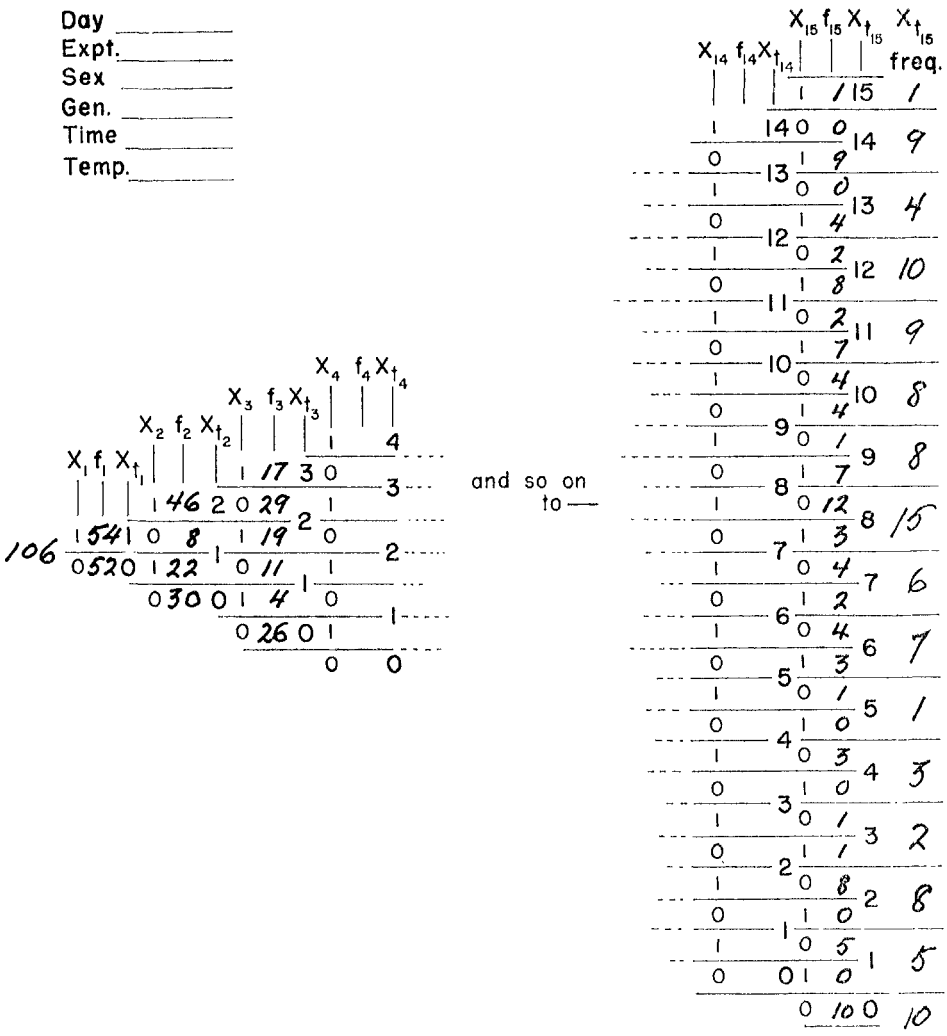


FIG. 1. MASS SCREENING SCORE FORM

combined for the next observation. Figure 1 shows the results schematically up through $X_{t_{15}}$.

The reason for the "experimental convenience" of dichotomous classes in the standard procedure should now be apparent; with more than two classes the number of subgroups becomes unmanageable.

5. *Analysis*

*a. The distribution of $X_t$ scores.* One of the objectives of experimental behavior genetics is reliable differentiation between individuals and subsequent genetic validation of differences by means of selective breeding. Since, for a given behavior, it is assumed that there is a range of ability and that the *S*s in a population are distributed over the range, it follows that any methods which tend to pile up the final scores in a few extreme categories should be eschewed in favor of others which distribute the scores more widely. The individuals

whose behavior is under observation will be used for breeding, hence it is important to differentiate them clearly on the behavioral scale. Failure to do this prevents the discovery of any genotypic differences that might exist.

The $E$ can usually control the form of the distribution of total $X_t$ scores. In our illustrative experiment this control was accomplished through selection of the time interval in which the response can be performed, i.e., the proportions $p$, of "passes" and $q$, of "fails" vary as a function of the amount of time allowed in the test tube. In examples from several experiments it may be shown that when $p > .5$, the $X_t$ distribution is negatively skewed and when $p < .5$, $X_t$ is positively skewed. Either type of skewness is undesirable because cases pile up in the extreme categories where, for the purposes of selective breeding, the finest differentiations are needed.

This point is illustrated in Table 1 where the frequency distribution of the composite score $X_{t_{10}}$ from Fig. 1 is presented in the first row of entries. A 15-sec. cutoff was used for this sample. The mean proportion earning a score of $X_1 = 1$ on the ten successive standard tests is $\bar{p} = .5$. The distribution is seen to be platykurtic with no appreciable piling up of the cases in the extreme categories. This is the result of the approximately 50–50 cut on each trial.

The effects of extreme cuts are shown in the other rows of Table 1. For the group with an 8-sec. cutoff in the standard test the proportion getting into the upper tube is $\bar{p} = .16$, with the result that the composite $X_{t_{10}}$ scores are very positively skewed with a pile up of flies in the 0 category. The opposite extreme cut of 27 sec. gives a $\bar{p} = .66$, with a pile up at the high $X_{t_{10}}$ scores.

*b. Reliability of $X_t$ scores.* It is important that the composite $X_t$ score be reliable if $E$ is to use the differentiations between individuals as the basis for further experimental work on selective breeding, conditioning, or the investigation of the generality of behavior $X$. The reliability coefficient, $r_{tt}$, cannot be computed by the split-half method in the mass screening method because combining into a single group all $S$s with the same composite $X_t$ score loses the specific sample score history of each individual. The coefficient can be estimated accurately, however, from the variances of the composite $X_t$ score and of the individual test sample scores, as follows (3, Formula 12):

$$r_{tt} = \frac{n}{n-1}\left(1 - \frac{\Sigma V_i}{V_t}\right), \qquad [1]$$

where:

$n$ = number of standard test samples or replications.

$\sum V_i$ = sum of the variances ($\sigma_i^2$) of the $n$ test samples.

TABLE 1

DISTRIBUTION OF INDIVIDUALS IN COMPOSITE $X_t$ SCORE

(Entries are frequencies)

| $p$ | Cutoff | $X_{t_{10}}$ | | | | | | | | | | | $N$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| .50 | 15 sec. | 11 | 5 | 9 | 7 | 10 | 8 | 11 | 13 | 15 | 11 | 6 | 106 |
| .16 | 8 sec. | 54 | 13 | 8 | 8 | 7 | 5 | 4 | 2 | 3 | 0 | 0 | 104 |
| .66 | 27 sec. | 0 | 9 | 3 | 3 | 4 | 4 | 4 | 16 | 24 | 12 | 13 | 92 |

$V_t$ = variance of the final composite $X_t$ scores, i.e., $\sigma_t^2$.

When, as in the present case, the standard procedure gives a dichotomous cut, the variance, $V_i$, of any particular sample observation is:

$$V_i = pq, \qquad [2]$$

where:

$p$ = proportion of individuals above the cut in all subgroups
= mean score when, as in the example, those above the cut are scored 1, those below 0.
$q = 1 - p$.

The values of the reliability coefficients and of other constants for several *Drosophila* experiments are given in the third rows of Table 2. The first group is the one presented in Fig. 1, in which 15 sample observations were finally taken under conditions believed to produce optimum differentiation between individuals. It will be noted that, beginning with the fourth column of entries, after the first few "adjustment" trials the reliabilities progressively increased to .87 for the final composite score based on 15 sample observations.

The *E* naturally asks: are the successive sample observations strictly comparable measures of the property $X$, here the negative geotropic reaction? The additional constants of Table 2 give insight into this question.

If the individuals systematically improve or deteriorate in performance the mean score, $p_i$, and the variance, $V_i = pq$, of successive observations will both change. In the first and second rows of Table 2 we see that in our example $p_i$ and therefore $V_i$ both remain relatively constant.

If the individuals become either more reliably differentiated or less so as screening proceeds, then the reliability coefficient will not increase according to the "Spearman-Brown law" of increased reliability with the addition of comparable sample observations. Evidence on this point can be secured in two ways.

TABLE 2

RELIABILITY COEFFICIENTS AND OTHER CONSTANTS IN THE DROSOPHILA GEOTROPIC EXPERIMENTS

| Group | | SAMPLE OBSERVATION, $X_i$ | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $n$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 15 sec. $N=106$ | $p_i$ | .64 | .38 | .50 | .53 | .47 | .53 | .55 | .63 | .55 | .48 | .49 | .48 | .46 | .46 |
| | $V_i$ | .23 | .24 | .25 | .25 | .25 | .25 | .25 | .23 | .25 | .25 | .25 | .25 | .25 | .25 |
| | $r_{tt}$ | .62 | .49 | .60 | .70 | .75 | .78 | .80 | .82 | .82 | .83 | .83 | .84 | .86 | .87 |
| | $n_c$ | 23 | 60 | 49 | 42 | 39 | 39 | 39 | 39 | 42 | 42 | 44 | 47 | 44 | 44 |
| | $\bar{r}_{ij}$ | .45 | .24 | .28 | .31 | .33 | .33 | .33 | .33 | .31 | .31 | .30 | .29 | .30 | .30 |
| 8 sec. $N=104$ | $p_i$ | .15 | .20 | .11 | .18 | .16 | .18 | .17 | .18 | .21 | | | | | |
| | $V_i$ | .13 | .16 | .10 | .15 | .14 | .15 | .14 | .15 | .17 | | | | | |
| | $r_{tt}$ | .66 | .63 | .68 | .68 | .72 | .75 | .77 | .80 | .81 | | | | | |
| | $n_c$ | 20 | 34 | 35 | 47 | 44 | 47 | 44 | 44 | 44 | | | | | |
| | $\bar{r}_{ij}$ | .49 | .36 | .35 | .29 | .30 | .29 | .30 | .30 | .30 | | | | | |
| 27 sec. $N=92$ | $p_i$ | .83 | .67 | .71 | .71 | .67 | .71 | .62 | .57 | .57 | | | | | |
| | $V_i$ | .14 | .22 | .21 | .21 | .22 | .21 | .24 | .25 | .25 | | | | | |
| | $r_{tt}$ | −.04 | .15 | .51 | .59 | .69 | .72 | .72 | .76 | .78 | | | | | |
| | $n_c$ | | 298 | 76 | 67 | 44 | 44 | 57 | 57 | 54 | | | | | |
| | $\bar{r}_{ij}$ | −.02 | .06 | .20 | .22 | .27 | .27 | .25 | .25 | .26 | | | | | |

The first is to discover whether the mean correlation, $\bar{r}_{ij}$, between sample observations entering into the composite, $X_t$, changes for successive $X_t$ scores. From the familiar Spearman-Brown approximation (**3**, Formula 17), we note that the reliability coefficient, $r_{t_n t_n}$, for any composite $X_t$ based on $n$ samples is:

$$r_{t_n t_n} = \frac{n\bar{r}_{ij}}{1+(n-1)\bar{r}_{ij}}, \qquad [3]$$

whence, solving for $\bar{r}_{ij}$:

$$\bar{r}_{ij} = \frac{r_{t_n t_n}}{n-(n-1)r_{t_n t_n}} \qquad [4]$$

The successive values of $\bar{r}_{ij}$ are given in Table 2, fifth rows. We note that after the first few trials $\bar{r}_{ij}$ plateaus around .30.

The other way is for $E$ to set a desired reliability for the final composite, and solve for the value of $n$ in Equation 3 that will achieve this desired reliability. Suppose $E$ desires a reliability of .95. Call this $R_{tt}$. Set $R_{tt}$ into Equation 3 and solve for $n$:

$$n = \frac{R_{tt}(1-\bar{r}_{ij})}{\bar{r}_{ij}(1-R_{tt})}. \qquad [5]$$

The values of $n$ for $R_{tt}=.95$ are given in the fourth rows of Table 2. In general they remain around 45 trials. This finding has the practical value of informing $E$ how many sample trials are necessary to achieve the reliability he desires. If $n$ turns out to be too large, a design having more classes per trial might be considered as a means of reducing the number of trials required.

When the individual test sample scores are not available, as is the case when groups are screened on the multiple-unit discrimination maze (**2**), the reliability coefficient can be computed directly from the final distribution of $X_t$ scores by means of the Total Score formula (**3**, Formula 37):

$$r_{tt} = \frac{n}{n-1}\left(1-\frac{M_t - M_t^2/n}{V_t}\right), \qquad [6]$$

where $M_t$ equals the mean of the final composite $X_t$ scores.

*c. Domain validity coefficient of the composite score, $X_t$.* The reliability coefficient, $r_{tt}$, though necessary in the above formulations, is not the best statement of the reliability of the composite $X_t$. A more meaningful index is the correlation between the $X_t$ scores and that on an indefinitely large number of screenings, namely $X_{t\infty}$. Though the "true score," $X_{t\infty}$ is not available, the correlation $r_{tt_\infty}$ can nevertheless be estimated as follows (**3**, Formula 21):

$$r_{tt_\infty} = \sqrt{r_{tt}}. \qquad [7]$$

Thus, in our case our $X_t$ based on fifteen screenings would correlate $r_{tt_\infty} = \sqrt{.867} = .93$ with a perfectly reliable measure based on many such screenings. This coefficient also has the following added meaning: If we had the true score of each fly based on many sets of 15 screenings, the ratio of the standard deviation of these true scores to that of the observed $X_t$ score would be .93. In short, the distribution of true scores would look much like that actually observed.

*d. Individual variance ("errors of measurement").* In order to conduct experiments on selective breeding, conditioning, or generality it is necessary to get a practical estimate of the amount of difference in $X_t$ scores among individuals that is undetermined, i.e., not assignable to known sources of variation. This estimate is the individual variance, $V_o$ (**3**, Formula 23a), where:

$$V_0 = V_t(1 - r_{tt}). \quad [8]$$

In our example for $X_{t_{15}}$, $V_t = 4.40$, hence the individual standard deviation is:

$$\sigma_0 = 4.40\sqrt{1 - .867} = 1.6.$$

The necessity of a high reliability can be seen in the above formula: as the reliability approaches unity the amount of variation attributable to individual variance tends to vanish.

*Nonuniformity of individual variance.* The individual variation, however, is most likely not constant over the final distribution: (*a*), an extreme score can vary in only one direction, towards the mean: (*b*), the individuals receiving extreme scores have shown perfectly consistent performance throughout, that is, either they have always scored a zero or they have always scored one. Hence, it might be expected that the individual variation, as estimated by a retest, should be much smaller at the extremes than in the middle of the distribution.

*Empirical check.* To assess this possibility a retest or validation experiment may be performed. In our illustration, the *S*s receiving extreme $X_t$ scores of 15 and 14 were combined and put through $n' = 10$ additional trials; also those receiving middle $X_t$ scores of 7 and 8 were put through a retest of 10 trials. For the extreme categories $\sigma_{t_{n'}}{}^2 = 4.00$, while for the middle categories $\sigma_{t_{n'}}{}^2 = 5.82$, the latter being significantly larger than the predicted variance for the middle categories. It is evident that the assumption of uniformity of individual variance over the whole $X_t$ scale is doubtful.

## Limits of Selective Breeding

How many generations is it necessary or practical to continue a selective breeding program, i.e., what are the criteria for stopping? The individual standard deviation, $\sigma_o = \sqrt{V_o}$, provides an answer to this question: it is useless to attempt further selection in any line beyond the point where its $\sigma_t = \sigma_o$; at that point the method of observation no longer reliably differentiates individuals, i.e., neither selection nor the evaluation of the results of selection are any longer possible. In our case, no further selective breeding would be attempted in any line whose $\sigma_t$ was much below 1.6.

## Summary

Fast breeding, prolific, small organisms are pre-eminently suited for studies in the field of behavior genetics. Their value as experimental *S*s is further enhanced by the method of mass screening that succeeds in combining the objective of reliable individual measurement with that of mass observation. Hence, it is now possible to achieve the experimental desiderata of efficiency, reliability, and brevity in the field of behavior genetics. The method is illustrated by experiments on the geotropic responses of *Drosophila*.

## REFERENCES

1. Guilford, J. P. *Psychometric methods.* (2nd Ed.) New York: McGraw-Hill, 1954.
2. Hirsch, J. A multiple-unit discrimination maze for the reliable mass screening of small organisms. *J. comp. physiol. Psychol.*, in press.
3. Tryon, R. C. Reliability and domain validity: reformulation and historical critique. *Psychol. Bull.*, in press.

*Received May 8, 1956.*