C. R. HENDERSON
Cornell University

Chapter 22

Specific and General Combining Ability

By general combining ability we mean the average merit with respect to some trait or weighted combination of traits of an indefinitely large number of progeny of an individual or line when mated with a random sample from some specified population. The merit of the progeny is measured in some specified set of environmental circumstances. If maternal effects are present, we must specify that the tested individuals are males. If the tested individuals are females, the merit of the progeny is a function of both general combining ability and maternal ability.

General combining ability has no meaning unless its value is considered in relationship to at least one other individual or line and unless the tester population and the environment are specified. For example, suppose two dairy bulls used concurrently in an artificial breeding ring each have 500 tested daughters, and that it can be assumed that the cows to which the two bulls were mated were a random sample of cows from herds using artificial breeding. Suppose that the mean of the butterfat records of the daughters of the first sire is 410 pounds and of the second sire is 400 pounds. Five hundred tested daughters are sufficient to reduce the sampling variance of the progeny mean to a negligible amount. Consequently the general combining ability of the first sire is 410 − 400 = 10 pounds better than that of the second in this particular population and in this set of environmental circumstances. The general combining abilities of the two sires might differ by more or by less than 10 pounds if they were used in some other region where both the genotypes of the cows to which they were mated as well as the environment could be quite different from those of the test.

SPECIFIC COMBINING ABILITY

We shall define specific combining ability as the deviation of the average of an indefinitely large number of progeny of two individuals or lines from

the values which would be expected on the basis of the known general combining abilities of these two lines or individuals and the maternal ability of the female parent. As Lush (1948) has pointed out, apparent specific effects, or what animal breeders usually call *nicking*, also can be a consequence of Mendelian sampling, of inaccurate estimates of the additive genetic values of the two parents, and of environments affecting the progeny which are different from the average environments in which the general combining abilities and the maternal abilities were estimated.

Genetically, specific combining ability is a consequence of intra-allelic gene interaction (dominance) and inter-allelic gene interaction (epistasis). We shall assume in this paper that we can estimate only the joint effect of dominance and epistasis. As an illustration of specific combining ability let us suppose that we know that the general combining ability with respect to weight in swine line A is $+10$ pounds at 154 days, and that the general combining ability plus maternal ability of line B is $+5$ pounds at 154 days. Then if an indefinitely large number of progeny of the cross $A \times B$ has a mean of $+7$ pounds, the specific effect for this cross is $7 - 10 - 5 = -8$.

SELECTION FOR GENERAL AND SPECIFIC COMBINING ABILITY

Under some circumstances selection would be largely for general combining ability, and in other circumstances for a combination of general and specific combining ability. For example, those selecting sires for use in a large artificial breeding ring are interested primarily in obtaining sires with the highest general combining ability with respect to the population of cows and environments in which the bulls are to be used. On the other hand, those wishing to employ crosses among inbred lines for commercial use select for a combination of general, maternal, and specific effects.

Now let us consider some of the problems involved in selecting for general and specific combining ability. There are reasonably good solutions to some of these problems, but almost none for others. Some of the questions which are involved are:

1. Given a particular set of records how can one best estimate the general combining abilities of individuals, families, or lines, and how can one best estimate the value of the progeny of a specific cross between families or inbred lines?

2. What proportion of the breeder's resources should be put into a testing program? For example, if he is dealing with inbred lines, what proportion of his resources should be employed in the making of lines and what proportion in testing them for general and specific combining ability?

3. Having decided on the size of the testing program, what kind of tests should be made? For example, should lines be tested in topcrosses or in line crosses or in some combination of these two procedures? Also what use should

be made of a sequential type of testing in which some lines are discarded on the basis of a very preliminary and inaccurate test?

4. What relative emphasis in selection should be placed on general as compared to specific combining ability?

5. How much inbreeding should be done in the making of lines? How fast should the lines be made?

Obviously a complete discussion of all these problems and their possible solutions in the time at our disposal is impossible. Consequently we shall discuss primarily the problem of estimating general combining abilities of lines and individuals and of estimating the values of specific crosses among lines, given a particular set of records. In addition, since estimates of the variances play an important role in these selection methods, we shall discuss briefly the problem of estimating variance components from the results of line-cross tests.

So far as estimation of general combining abilities of individuals is concerned, the methods to be presented here are essentially those of the selection index. It will be shown that no assumption of normality of distributions is required; that joint estimates of general combining abilities and certain parameters such as the population means, the yearly effect, the age and inbreeding effect, can be obtained; and that certain short-cut computational procedures are sometimes distinctly advantageous. An application of the principles of the selection index to estimation of general combining abilities of lines or families also will be presented. Finally it will be shown that application of the selection index need not be restricted as it has been to selection for additive effects, but can be applied equally well to joint selection for specific effects and general combining ability. The selection index approach to appraising crosses can, under some circumstances, be much more efficient than selection based on the mean of the progeny of a particular cross.

ESTIMATION PROBLEMS IN SELECTION

Before turning to selection for general and specific combining abilities let us consider the type of estimation problem which is involved and some general solutions to it. Later the manner in which the solutions can be applied to our present problem will be discussed. Our estimation problem can be stated in this way. We have a sample of N observations, $y_1, y_2, \ldots, y_N$, from which we wish to estimate $\theta_1, \theta_2, \ldots, \theta_q$. The y's are assumed to have a multivariate distribution (precisely what distribution need not be specified for the present) with means, $b_1x_{1i} + b_2x_{2i} + \ldots + b_px_{pi}$, and variance-covariance matrix,

$$\| \sigma_{y_i y_j} \|.$$

The b's are fixed parameters such as the population mean and the regression of y on age of the dam, and x is an observable parameter, the first subscript denoting with which b it is associated, the second subscript with

which sample observation. As an illustration, x_1 might be associated with b_1, the population mean. Then x_{1i} would have the value 1 in each observation; x_{2i} might denote the inbreeding coefficient of the dam.

Now comes the really crucial part of the model. The θ's are regarded as having some multivariate distribution with means zero and variance-covariance matrix,

$$\| \sigma_{\theta_g \theta_h} \|.$$

Also the θ's and y's are regarded as having a joint distribution with covariances $\sigma_{\theta_h y_i}$. The way in which this problem differs from the ordinary estimation problem in statistics is that here we wish to estimate the values of individual θ's which are regarded as a sample from some specified population.

Selection for Additive Effects in the Normal Distribution

What is the "best" way to estimate the θ's? Suppose that they represent additive genetic values of individuals and that any linear function of the y's is normally distributed. Lush (1948) has shown that, subject to the normality assumptions, improvement in additive genetic merit of a population through selection by truncation of the estimates (indexes) of additive genetic values is maximized by choosing that index which has maximum correlation with additive genetic value. This principle has been used in the index method of selection by Fairfield Smith (1936), Hazel (1943), and others. These workers have shown that the index can be found in a straightforward manner provided certain variances and covariances and all of the b's, the fixed elements of the model, are known.

The values of $K_{\theta i}$ which maximize $r_{\theta\hat{\theta}}$ where $\hat{\theta} = K_{\theta 1} w_1 + \ldots + K_{\theta N} w_N$ are the solution to the set of simultaneous equations (1). The w's are the y's corrected for the fixed elements of the model such as the population mean (not the sample mean). Thus $w_1 = y_1 - b_1 x_{11} - \ldots - b_p x_{p1}$.

$$\begin{aligned}
K_{\theta 1}\sigma^2_{y_1} + K_{\theta 2}\sigma_{y_1 y_2} + \ldots + K_{\theta N}\sigma_{y_1 y_N} &= \sigma_{y_1 \theta} \\
K_{\theta 1}\sigma_{y_1 y_2} + K_{\theta 2}\sigma^2_{y_2} + \ldots + K_{\theta N}\sigma_{y_2 y_N} &= \sigma_{y_2 \theta} \\
\cdot \qquad\qquad\qquad\qquad\qquad\qquad &\quad \cdot \\
\cdot \qquad\qquad\qquad\qquad\qquad\qquad &\quad \cdot \\
\cdot \qquad\qquad\qquad\qquad\qquad\qquad &\quad \cdot \\
K_{\theta 1}\sigma_{y_1 y_N} + K_{\theta 2}\sigma_{y_2 y_N} + \ldots + K_{\theta N}\sigma^2_{y_N} &= \sigma_{y_N \theta}
\end{aligned} \qquad (1)$$

Selection when Form of Distribution is Unspecified and *b*'s Are Unknown

Maximization of $r_{\theta\hat{\theta}}$ is a satisfactory solution to the problem of selection for additive genetic values under the normality assumption and the assumption of known b's. Is a comparable solution available when nothing is known of the distribution or of the b's? So far as I am aware there is not.

Consequently let us consider some other criterion of a "best" index. We shall use as our criterion of "best" that index from the class of linear functions of the sample which is unbiased (coefficients of all b's $= 0$ in $E\hat{\theta}$) and for which $E(\hat{\theta} - \theta)^2$ is a minimum. E denotes expected value. Consequently $E(\hat{\theta} - \theta)^2$ denotes the average in repeated sampling of the squared deviations of the index of θ about the true value of θ. When the b's are unknown, the same criterion of best is applied to them, that is, minimum $E(\hat{b} - b)^2$ for unbiased estimates ($E\hat{b} = b$) which are linear functions of the sample. It turns out that minimization of $E(\hat{\theta} - \theta)^2$ and maximization of $r_{\theta\hat{\theta}}$ lead to identical indexes. Hence the assumption of normality is not essential to construction of selection indexes as now used.

It must be obvious that the selection index method just described is very laborious when a number of different θ need to be estimated, for the solution to a set of simultaneous equations is required for each θ. In practice this difficulty is avoided to a certain extent by choosing arbitrarily only a few sources of information to be employed in selection. This is not a wholly satisfactory solution, for in most cases if the number of different indexes is not to be entirely too large, information must be rejected which could add at least a little to the accuracy of the index.

By means of a simple modification it becomes necessary to solve only one set of equations no matter how many θ are estimated from a particular set of data, and precisely the same index as in the conventional method is obtained. Using the same notation as before, the index for θ is now

$$\hat{\theta} = C_1\sigma_{\theta y_1} + C_2\sigma_{\theta y_2} + \ldots + C_N\sigma_{\theta y_N},$$

where the C's are the solution to a set of equations identical to set (1) except that the right members are $w_1, \ldots, w_N$ rather than $\sigma_{\theta y_1}, \ldots, \sigma_{\theta y_N}$. Consequently once the C's are computed, any number of θ's can be estimated simply by taking the appropriate linear function of the C's.

More tedious computations result if the b's are not known. One solution is of the following general form. In order that each θ be unbiased it is necessary that the K's have these restrictions imposed:

$$\begin{aligned} K_1x_{11} + K_2x_{12} + \ldots + K_Nx_{1N} &= 0 \\ K_1x_{21} + K_2x_{22} + \ldots + K_Nx_{1N} &= 0 \\ \cdot \qquad\qquad\qquad\qquad & \cdot \\ \cdot \qquad\qquad\qquad\qquad & \cdot \\ K_1x_{p1} + K_2x_{p2} + \ldots + K_Nx_{pN} &= 0 \end{aligned} \qquad (2)$$

Subject to these restrictions the values of the K's which minimize $E(\hat{\theta} - \theta)^2$ can then be found.

If we wish to obtain estimates of the b's which are unbiased and have minimum $E(\hat{b} - b)^2$, we impose the restrictions of equations (2) except that

the right member of the equation pertaining to the particular b to be estimated is 1 rather than 0.

An easier solution to the problem of unknown b's often can be obtained by regarding the model as,

$$y_i = b_1 x_{1i} + b_2 x_{2i} + \theta_1 z_{1i} + \ldots + \theta_q z_{qi} + e_i ,$$

where the e_i are independently distributed with mean zero and variance $\sigma^2_{e_i}$ and the z's are observable parameters. For example, θ_1 might represent the general combining ability of inbred line A, θ_2 the general combining ability of line B, and θ_3 a specific effect peculiar to the cross $A \times B$. The observable parameters z would have the following values: $z_1 = 1$ when line A is one of the parents, $= 0$ otherwise; $z_2 = 1$ when line B is one of the parents, $= 0$ otherwise; and $z_3 = 1$ when y_i is an observation on the cross $A \times B$ or $B \times A$, $= 0$ otherwise. Now the joint estimates of b's and θ's are the joint solution to the subsets of equations (3), (4), and (5).

$$\begin{aligned} C_1\sigma^2_{y_1} + C_2\sigma_{y_1y_2} + \ldots + C_N\sigma_{y_1y_N} &= y_1 - \hat{b}_1 x_{11} - \ldots - \hat{b}_p x_{p1} \\ &\vdots \\ C_1\sigma_{y_1y_N} + C_2\sigma_{y_2y_N} + \ldots + C_N\sigma^2_{y_N} &= y_N - \hat{b}_1 x_{1N} - \ldots - \hat{b}_p x_{pN} \end{aligned} \tag{3}$$

$$\begin{aligned} \hat{\theta}_1 &= C_1\sigma_{y_1\theta_1} + \ldots + C_N\sigma_{y_N\theta_1} \\ &\vdots \\ \hat{\theta}_q &= C_1\sigma_{y_1\theta_q} + \ldots + C_N\sigma_{y_N\theta_q} \end{aligned} \tag{4}$$

$$\begin{aligned} \hat{b}_1 Sx_1^2 + \ldots + \hat{b}_p Sx_1x_p + \hat{\theta}_1 Sx_1z_1 + \ldots + \hat{\theta}_q Sx_1z_q &= Sx_1y \\ &\vdots \\ \hat{b}_1 Sx_1x_p + \ldots + \hat{b}_p Sx_p^2 + \hat{\theta}_1 Sx_pz_1 + \ldots + \hat{\theta}_q Sx_pz_q &= Sx_py , \end{aligned} \tag{5}$$

where

$$Sx_1^2 = \sum_i \frac{x_{1i}^2}{\sigma^2_{e_i}}, \qquad Sx_1x_2 = \sum_i \frac{x_{1i}x_{2i}}{\sigma^2_{e_i}}, \quad \text{etc.}$$

These equations can be solved by the following steps. First solve for the C's in equations (3). The results will be in terms of the sample observations and the $\hat{b}$'s. Second, substitute values of these C's in equations (4) to obtain $\hat{\theta}$'s in terms of the sample and the $\hat{b}$'s. Third, substitute these values of the $\hat{\theta}$'s in equations (5) and solve for the $\hat{b}$'s. Fourth, substitute the computed values of the $\hat{b}$'s in (4) and solve for the $\hat{\theta}$'s.

An alternative computational procedure which is less laborious when the θ's are few in number, and in particular when the θ's are uncorrelated, involves joint estimation of the b's and θ's by solution of equations (5) to which are added equations (6).

$$\begin{aligned}
&b_1 Sx_1 z_1 + \ldots + b_p Sx_p z_1 + \theta_1 (S z_1^2 + \sigma^{11}) + \ldots + \theta_q (S z_1 z_q + \sigma^{1q}) = S z_1 y \\
&\cdot \qquad\qquad \cdot \qquad\qquad \cdot \qquad\qquad \cdot \\
&\cdot \qquad\qquad \cdot \qquad\qquad \cdot \qquad\qquad \cdot \qquad\qquad (6) \\
&\cdot \qquad\qquad \cdot \qquad\qquad \cdot \qquad\qquad \cdot \\
&b_1 Sx_1 z_q + \ldots + b_p Sx_p z_q + \theta_1 (S z_1 z_q + \sigma^{1q}) + \ldots + \theta_q (S z_q^2 + \sigma^{qq}) = S z_q y ,
\end{aligned}$$

where

$$\| \sigma^{ij} \| = \| \sigma_{\theta_i \theta_j} \|^{-1} ,$$

and

$$Sx_1 z_1 = \sum_i \frac{x_{1i} z_{1i}}{\sigma_{e_i}^2}, \quad \text{etc.}$$

These equations are simply least squares equations (the θ's are regarded as fixed rather than having a distribution) modified by adding σ^{ij} to certain coefficients.

SELECTION BY MAXIMUM LIKELIHOOD ESTIMATES

Now let us assume that the θ's have the multivariate normal distribution and that the errors are normally and independently distributed. What are the maximum likelihood estimates of the θ's and b's? It just so happens that the estimates which are unbiased and which have minimum $E(\hat{\theta} - \theta)^2$ and $E(\hat{b} - b)^2$ for the class of linear functions of the sample are also the maximum likelihood estimates. Consequently the estimation procedure we have described can be seen to have the following desirable properties: unbiasedness, maximum relative efficiency of all linear functions of the sample, maximization of genetic progress through selection by truncation when the distributions are normal, properties of maximum likelihood estimates when the distributions are normal, and equations of estimation which can be set up in a routine manner.

Unknown Variances and Covariances

An important problem in selection remains unsolved and perhaps there is no practical solution to it. What should be done if the variances and covariances are unknown? If our sample is so large that estimates of the variances and covariances can be obtained from it with negligible errors, we can use these estimates as the true values. Similarly we may be able to utilize estimates obtained in previous experiments. But if there are no data available other than a small sample, the only reasonable advice would seem to be to estimate the variances from the sample, perhaps modifying these estimates

somewhat if they appear totally unreasonable. At any rate the estimation procedure serves to point out what additional information is needed if an intelligent job of selection is to be accomplished.

SELECTION FOR ADDITIVE GENETIC VALUES IN INDIVIDUALS

As our first application of the methods described above, consider the estimation of additive genetic values of individuals with respect to a single trait (the single trait might be net merit) from a set of records all made in the same herd or flock. It will be assumed for the present that the population mean is known and that records can be corrected satisfactorily for all nonrandom environmental factors. For example, the records might represent all of the 305 day, mature equivalent butterfat records made in a herd during the past ten years. It is desired on the basis of these records to decide which cows should be culled, which heifers should be selected for replacements, and which bull calves should be grown out for possible use as herd sires.

In the usual approach to this selection problem by use of the selection index, one would decide what particular subset of the records would contribute most to the estimate of the value of each animal under consideration and would then construct separate indexes. The method to be presented here employs all available records in estimating the value of each animal. That is, no prior decision is made concerning which records to use to construct the index for each animal, but instead all available ones are used.

The first step in the procedure is the computation of what Emik and Terrill (1949) have called a numerator relationship chart and Lush (1948) has called genic variances and covariances for all animals whose records are to be used in the index or whose breeding values are to be estimated. In terms of Wright's (1922) coefficients of relationship and inbreeding, the genic variance of the ith animal is $1 + F_i$, where F_i is the inbreeding coefficient of the ith animal, and the genic covariance between the ith and jth animal is

$$R_{ij}\sqrt{(1+F_i)(1+F_j)},$$

where R_{ij} is the coefficient of relationship between the two animals. The numerator relationship or genic covariance, which we shall denote by a_{ij}, is the numerator of the fraction representing relationship. That is

$$R_{ij} = \frac{a_{ij}}{\sqrt{(1+F_i)(1+F_j)}}.$$

The computation of $\|a_{ij}\|$ is a routine procedure if it is done systematically as described by Emik and Terrill and by Lush.

Next we need an estimate of heritability of the trait, and if more than one record is available on a single animal, as would be true of butterfat production, an estimate of repeatability. Now let $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p$ be the mean of the n_i records of each of p animals, these records having been corrected for

non-random environment and expressed as deviations about the population mean. The next step is to solve the following set of equations for $C_1, \ldots, C_p$. In these equations h denotes heritability and r denotes repeatability.

$$\begin{aligned}
&C_1\left(F_1h+\frac{1+(n_1-1)\,r}{n_1}\right)+C_2a_{12}h+\ldots+C_pa_{1p}h=\bar{y}_1\\
&C_1a_{12}h+C_2\left(F_2h+\frac{1+(n_2-1)\,r}{n_2}\right)+\ldots+C_pa_{2p}h=\bar{y}_2\\
&\quad\cdot\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\cdot\\
&\quad\cdot\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\cdot\\
&\quad\cdot\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\cdot\\
&C_1a_{1p}h+C_2a_{2p}h+\ldots+C_p\left(F_p+\frac{1+(n_p-1)\,r}{n_p}\right)=\bar{y}_p
\end{aligned}\tag{7}$$

If all available records are to be used in the estimation procedure just described, the number of equations to be solved for the C's is large. It might appear, in fact, that the number is too great for the method to have any value. However, the equations are ideally suited to an iterative solution. The reason for this is that the diagonal elements of the left members of the equations are very large compared to the off-diagonal elements thereby making the iterative solution a particularly rapid one. On the basis of our experience with a few herds a solution to sufficient accuracy can be obtained in three or four rounds of iteration.

Once the C's have been computed the estimate of g_i, additive genetic value of the ith animal, is

$$\hat{g}_i = h\,(C_1a_{1i}+C_2a_{2i}+\ldots+C_pa_{pi})\,.$$

If the ith animal had one or more records included in the computation of the C's the estimate can be computed more easily, for

$$\hat{g}_i = \bar{y}_i - C_i\,\frac{1+(n_i-1)\,r-n_ih}{n_i}\,.$$

The estimate of the real producing ability of a tested animal is even more simple to express. The estimated real producing ability is

$$\bar{y}_i - C_i\,\frac{(1-r)}{n_i}\,.$$

It should be pointed out that this estimate differs from the one presented by Lush (1945) since his method does not utilize records on relatives.

Valuable characteristics of the method just described, in addition to its ease of computation and its use of all available information, is that the inclusion of the records of the contemporaries of the ancestors of the animals being appraised automatically eliminates the troublesome problem of what effect selection has had on the phenotypic and genetic variances of the selected group of ancestors. Also changes in additive genetic variances and covari-

ances effected by inbreeding are automatically taken into account. If selection is intense, the sample mean may considerably overestimate the population mean appropriate for subtraction from the records. The safest procedure is to regard μ as unknown and to estimate it by the procedure described earlier (equations 3, 4, 5). It is also of interest to note that joint estimation by this method of such factors as environmental trends and age effects automatically eliminates biases in the estimates resulting from use of selected data.

SELECTION FOR GENERAL COMBINING ABILITY IN TOPCROSS TESTS

When it comes to estimation of the general combining abilities of inbred lines or of the values of specific crosses, apparently no application has been made of the selection index method. This failure may have been due to difficulty in obtaining the estimates of the needed variances and covariances, failure to see that the method was applicable, or the opinion that since inbred lines can be carefully tested more efficient but complex methods of appraisal are not worth the extra computational labor. We propose to show here how the methods can be applied to such selection problems, to indicate some situations in which it may result in considerably more efficiency in selection than the use of the straight means of the lines or crosses as the criteria of selection, and to present some approximate solutions which are relatively easy to compute.

Let us consider first one of the most simple tests of lines, the topcross test. In this test a random sample of individuals from each of several lines is mated to a tester population, and measurements are taken on the resulting progeny. If only one trait is considered important, the lines are usually rated according to the means of their topcross progeny. This method of ranking is as good as any, provided either that the same number of progeny is obtained for each line or that the sampling errors of the line means are negligible. Seldom, at least in large animal tests, would either of these conditions hold. Accidents usually preclude attainment of equal numbers, and sampling errors are usually large. If sequential testing is done, numbers would always be unequal. By sequential testing we mean here that lines are given a preliminary test, and a certain fraction of those performing worst are discarded. Then the remaining lines, accompanied perhaps by some new lines, are given another test, and so on through any number of cycles desired. The lines surviving several such tests would obviously have larger numbers of progeny than the new lines, and it would be a very inefficient procedure to disregard the results of prior tests on the older lines when choosing between them and the newer, less well-tested lines.

The way in which the lines should be ranked on the basis of all information is analogous to choosing between individuals with different numbers of records. In the latter case both repeatability of single records and the number of

records need to be considered; in the former case the genetic differences among lines, the environmental variance, and the number of progeny. Also in both cases consideration of the genetic covariances between individuals or between lines increases the accuracy of the ranking.

Assuming that the population mean is known and that it and non-random environmental factors have been subtracted from the means of the progeny of the various lines, the estimate of g_i, the general combining ability of the ith line, is

$$\hat{g}_i = C_1\sigma_{g_i\bar{y}_1} + \ldots + C_p\sigma_{g_i\bar{y}_p}$$

where $\bar{y}_1, \ldots, \bar{y}_p$ are the corrected means for the p tested lines and the C's are the solution to a set of equations with

$$\| \sigma_{\bar{y}_i\bar{y}_j} \|$$

as coefficients in the left members and corrected $\bar{y}_1, \ldots, \bar{y}_p$ as the right members. Computation of $\sigma_{\bar{y}_i\bar{y}_i}$ and $\sigma_{g_i\bar{y}_i}$ requires good estimates of

$$\| \sigma_{g_ig_j} \|$$

and of σ_e^2. Assuming that the corrected mean of a particular topcross is $\bar{y}_i = g_i + \bar{e}_i$, and that the errors are independent with common variance σ_e^2 we have the following variances and covariances

$$\sigma^2_{\bar{y}_i} = \sigma^2_{g_i} + \sigma^2_e/n_i \qquad \sigma_{\bar{y}_ig_i} = \sigma^2_{g_i}$$

$$\sigma_{\bar{y}_i\bar{y}_j} = \sigma_{g_ig_j} \quad \text{(where } i \neq j\text{)} \qquad \sigma_{\bar{y}_ig_j} = \sigma_{g_ig_j} \quad (i \neq j)$$

Frequently good estimates of μ and non-random environmental factors are not available and consequently must be estimated from the topcross data. For example, it is very likely that the environment is not the same from test to test and must be taken into account if the data from several tests are to be combined into a "best" index. In such cases the method of equations (5) and (6) can be employed to distinct advantage unless

$$\| \sigma_{g_ig_j} \|^{-1}$$

is too difficult to compute. To illustrate this method as applied to topcross data we shall assume that y_{ij}, the record of the jth progeny of the ith line, can be represented by

$$y_{ij} = b_1x_{1ij} + b_2x_{2ij} + g_i + e_{ij}\,.$$

b_1 and b_2 are examples of fixed parameters, g_i is the general combining ability of the ith line, and e_{ij} is a random error. Assuming that the g_i are distributed with means zero and known variance-covariance matrix,

$$\| \sigma_{g_ig_j} \|\,,$$

and that the e_{ij} are independently distributed with means zero and common variance σ_e^2, the estimates of the b's and g's which are "best" by the criterion used in this paper are the solution to the following equations:

$$\begin{aligned}
&\hat{b}_1 \sum_i \sum_j x_{1ij}^2 + \hat{b}_2 \sum_i \sum_j x_{1ij}x_{2ij} + \sum_i \hat{g}_i x_{1i.} = \sum_i \sum_j x_{1ij}y_{ij} \\
&\hat{b}_1 \sum_i \sum_j x_{1ij}x_{2ij} + \hat{b}_2 \sum_i \sum_j x_{2ij}^2 + \sum_i \hat{g}_i x_{2i.} = \sum_i \sum_j x_{2ij}y_{ij} \\
&\hat{b}_1 x_{11.} + \hat{b}_2 x_{21.} + \hat{g}_1 (n_1 + \sigma_e^2 \sigma^{11}) + \sum_{i \neq 1} \hat{g}_i \sigma_e^2 \sigma^{1i} = y_{1.} \qquad (8) \\
&\cdot \qquad\qquad\qquad\qquad\qquad\qquad \cdot \\
&\cdot \qquad\qquad\qquad\qquad\qquad\qquad \cdot \\
&\cdot \qquad\qquad\qquad\qquad\qquad\qquad \cdot \\
&\hat{b}_1 x_{1p.} + \hat{b}_2 x_{2p.} + \hat{g}_p (n_p + \sigma_e^2 \sigma^{pp}) + \sum_{i \neq p} \hat{g}_p \sigma_e^2 \sigma^{pi} = y_{p.}
\end{aligned}$$

Dots in the subscripts denote summation over that subscript, and σ^{ij} denotes an element of

$$\| \sigma_{g_i g_j} \|^{-1} .$$

The above procedure for appraising lines on the basis of topcrosses assumes either that the lines are homozygous or that only one progeny is obtained from each randomly chosen male. If these assumptions are not correct, the procedure is modified to take into account intra-line variances and covariances and the number of progeny per male.

What are the consequences of appraising lines on the basis of the arithmetic average of their respective progeny as compared to the more efficient method just described? First, the errors are larger than necessary. Second, selection of some small fraction of tested lines will tend to include a disproportionately large number of the less well-tested lines. The more efficient method discounts the higher averages in accordance with the number of tested progeny and the relative magnitudes of σ_g^2 and σ_e^2.

What if the number of lines tested is large and certain lines are related? This means that a large matrix,

$$\| \sigma_{g_i g_j} \| ,$$

has to be inverted and then a large set of simultaneous equations solved. What approximations might be employed in the interest of reducing computations? For one thing, we might ignore the covariances between the g's, thereby reducing the inverse matrix to $1/\sigma_{g_i}^2$ in the diagonal elements and 0 in the off-diagonal elements. Also if we know μ and non-random environmen-

tal factors well enough, further simplification is possible. Let $\bar{w}_i$ be the corrected mean of the progeny of ith line. Then

$$\hat{g}_i = \frac{n_i \sigma^2_{g_i}}{n_i \sigma^2_{g_i} + \sigma^2_e} \bar{w}_i .$$

This result is a straightforward application of the principles of the selection index.

It must be quite apparent that efficient appraisals of the general combining abilities of lines depend on knowledge of the variances and covariances of general combining abilities and of the variance of error. It hardly seems likely that estimates of the line variances and covariances can be obtained with accuracy comparable to estimates of additive genetic variances and covariances with respect to individuals. The latter estimates are based on studies of heritability and on the known facts of the hereditary mechanism. In the case of inbred lines, however, the sample of different lines tested is usually so small as to make the estimates of σ^2_g less reliable than we should like. A way around this difficulty in the case of traits for which heritability is well known is to compute the expected variances and covariances based on knowledge of σ^2_g in the original population from which the lines were formed, the inbreeding of the different lines, and the relationships between pairs of lines. It seems likely that such estimates would be more reliable when the number of lines is small than would estimates arising from the actual line tests. We cannot be any more precise regarding this point until methods are developed for placing confidence limits on estimates of variances and covariances arising from non-orthogonal data.

SELECTION FOR GENERAL COMBINING ABILITY, MATERNAL ABILITY, AND SPECIFIC ABILITY IN LINE CROSS TESTS

If we wish to estimate the general combining ability of lines relative to the population from which the lines themselves can reasonably be regarded as a random sample, line crosses give, for fixed size of testing facilities, more accurate estimates than do topcrosses. The reason for this is that we obtain from each cross estimates of the general combining abilities of two or more lines. Also, line crosses enable one to estimate differences in maternal abilities unconfounded with differences in general combining abilities and to appraise the values of specific crosses. In those species for which hand mating is the customary procedure, little more labor is required for line cross than for topcross tests. The estimation of line and line cross characteristics from line cross data is no different in principle from what we have already described with respect to estimation of additive genetic values of individuals or general combining abilities of lines. As before, we wish to obtain unbiased and most efficient estimates of certain genetic values. For the sake of simplicity of presentation we

shall confine ourselves to discussion of the analysis of single crosses. Application of these principles to multiple cross data involves no new principles.

Let us consider first what type of model might be reasonable for a single cross. It is not too difficult to suppose that the value of a particular observation on a single cross is the sum of the general combining ability of the male line, the general combining ability of the female line, a maternal effect coming from the line used as the female, a specific effect due to dominance and epistasis and peculiar to the particular cross, non-random environmental effects, and a multitude of random errors such as Mendelian sampling and the environment peculiar to the particular progeny on whom the record is taken. More complicated models could of course be proposed, but the one which we have just described would seem to account for the major sources of variation among crosses. Furthermore it is amenable to mathematical treatment. Putting the above description in a mathematical model we have

$$y_{ijk} = b_1 x_{1ijk} + b_2 x_{2ijk} + g_i + g_j + m_j + s_{ij} + e_{ijk} ,$$

where y_{ijk} is the observation on the kth progeny of a cross between the ith line used as a male parent and the jth line as a female parent, the b's and x's are related to the mean and other non-random environmental factors as described in the model for the topcross test, $g_i(g_j)$ is the general combining ability of the ith(jth) line, m_j is an effect in addition to the additive genetic value which is common to all progeny of the jth line used as a female parent, s_{ij} is an effect over and above the additive genetic and maternal effects and which is common to all progeny of the cross of the ith line by the jth line or of the jth line by the ith line, and e_{ijk} is a random error associated with the particular observation.

In this model the g_i are regarded as having some multivariate distribution with means zero and variance-covariance matrix,

$$\| \sigma_{g_i g_j} \| .$$

The m_j, s_{ij}, and e_{ij} are all regarded as independently distributed with means zero and variances σ_m^2, σ_s^2, and σ_e^2, respectively. It is of course conceivable that the variances of the m_j and s_{ij} and the covariances between them vary with the inbreeding and relationships of the lines. Also g_i and m_j may be correlated. In the absence of any real knowledge concerning such covariances we shall ignore them for our present purposes. If, however, something is known about these covariances, the estimation procedure can be modified to take them into account. The procedure should also be modified if the lines are not homozygous and each parent has more than one progeny.

A single cross test can supply answers to the following questions with respect to the lines tested:

1. What are the best estimates of the relative values of the tested lines

when used as the male parent in topcrosses on the population from which the lines are regarded as a sample?

2. What are the best estimates of the relative values of the tested lines as female parents in crosses with males from the above population?

3. What are the best estimates of the relative values of specific single crosses among the tested lines?

Suppose that n_{ij} progeny of the cross ith line of male by jth line of female are tested (n_{ij} can be zero for some crosses). Now the easiest way to estimate the value of the ith line as a male parent is simply to compute the mean of all progeny of the line when used as the male parent. This simple procedure, however, fails to take into account the distribution among lines of the mates of males of the ith line, the covariances among the general combining abilities of lines, the consequences of specific effects, the size of the error variance, and the number of progeny tested. Furthermore, since the ith line is used also as the female parent in certain crosses, something can be gained by employing the measurements on these progeny. Estimation by the general procedure we have described takes into account all of these factors. Similarly the easiest way to estimate the maternal ability of the jth line is to compute the mean of all progeny out of females of the jth line, but the most efficient procedure takes into account the same factors as are needed in efficient estimation of general combining ability. Finally the easy way to appraise the value of a particular cross is merely to find the mean of all progeny of the specific cross (if that cross has been tested). This latter estimate is subject to large sampling error since it would seldom be feasible to test many individuals of the numerous possible crosses among even a few lines. The error of estimation can be materially reduced by utilizing the fact that the true merit of a cross is a function of the general combining abilities of two lines, the maternal ability of the female line, and the specific effect peculiar to that cross and to its reciprocal. The method to be described places the proper emphasis on estimates of general and maternal abilities and on the progeny averages of the specific cross and its reciprocal. The procedure also enables estimates to be made of the value of a specific cross even though that particular cross has not been tested.

The major step in these efficient estimation procedures is the setting up and solving of a set of simultaneous equations in the $\hat{b}$'s, $\hat{g}$'s, $\hat{m}$'s, and $\hat{s}$'s. These equations are as follows:

$$\hat{b}_1 \sum_i \sum_j \sum_k x_{1ijk}^2 + \hat{b}_2 \sum_i \sum_j \sum_k x_{1ijk}x_{2ijk} + \sum_i \hat{g}_i (x_{1i..} + x_{1.i.}) + \sum_j \hat{m}_j x_{1ij.} + \sum_{i<j} \hat{s}_{ij} (x_{1ij.} + x_{1ji.}) = \sum_i \sum_j \sum_k x_{1ijk} y_{ijk} \tag{9}$$

and similarly for the b_2 equation.

$$\hat{b}_1(x_{11..}+x_{1.1.})+\hat{b}_2(x_{21..}+x_{2.1.})+\hat{g}_1(n_{1.}+n_{.1}+\sigma_e^2\sigma^{11})$$
$$+\sum_{i\neq 1}\hat{g}_i(n_{i1}+n_{1i}+\sigma_e^2\sigma^{1i})+\hat{m}_1 n_{.1}+\sum_{j\neq 1}m_j n_{1j}+\sum_{j\neq 1}\hat{s}_{1j}(n_{1j}+n_{j1})$$
$$=y_{1..}+y_{.1.}$$

and similarly for the other g_i equations.

$$\hat{b}_1 x_{1.1.}+\hat{b}_2 x_{2.1.}+\hat{g}_1 n_{.1}+\sum_{i\neq 1}\hat{g}_i n_{i1}+\hat{m}_1(n_{.1}+\sigma_e^2/\sigma_m^2)+\sum_{i\neq 1}\hat{s}_{1i}n_{i1}=y_{.1.}$$

and similarly for the other m_j equations.

$$\hat{b}_1(x_{112.}+x_{121.})+\hat{b}_2(x_{212.}+x_{221.})+(\hat{g}_1+\hat{g}_2)(n_{12}+n_{21})+\hat{m}_1 n_{21}$$
$$+\hat{m}_2 n_{12}+\hat{s}_{12}(n_{12}+n_{21}+\sigma_e^2/\sigma_s^2)=y_{12.}+y_{21.}$$

and similarly for the other s_{ij} equations.

These equations are not particularly difficult to solve, for each s_{ij} can be expressed as a function of $y_{ij.}$, $y_{ji.}$, $\hat{b}_1, \hat{b}_2, \hat{g}_i, \hat{g}_j$, and $\hat{m}_j$. Utilizing this relationship the equations can be reduced to a set involving none of the $\hat{s}_{ij}$. Also an iterative solution is usually easy because of the relatively large diagonal coefficients. Once the estimates of g_i, m_j, and s_{ij} are obtained it is a simple matter to evaluate the lines and crosses. The estimate of the value of a line as the male parent in topcrosses is $\hat{g}_i$, and the estimate of its average value as the female parent is $\hat{g}_i+\hat{m}_i$. The value of a single cross is estimated simply as $\hat{g}_i+\hat{g}_j+\hat{m}_j+\hat{s}_{ij}$. It is appropriate to add the estimates in this manner because they have the desirable property of invariance.

If solution of the large set of simultaneous equations required for most efficient appraisal of lines is considered too burdensome, certain approximate solutions can be employed. An approximation suggested by the common practice in construction of selection indexes is the choosing of certain information most pertinent to the particular line or cross to be appraised. For example, the estimate of g_i might be based entirely on $y_{i..}$ and $y_{.i.}$, each corrected for the b's as best can be done with the information available regarding their values. As a further simplification it might be assumed that the g_i are uncorrelated and have common variance σ_g^2. Similarly m_i might be estimated entirely from $y_{i..}$ and $y_{.i.}$. These approximate solutions are

$$\hat{g}_i=C_1\sigma_{y_i y_{i..}}+C_2\sigma_{y_i y_{.i.}}$$

$$\hat{m}_i=C_1\sigma_{m_i y_{i..}}+C_2\sigma_{m_i y_{.i.}},$$

where the C's are the solution to

$$C_1\sigma^2_{y_{i..}}+C_2\sigma_{y_{i..}y_{.i.}}=y_{i..}-\hat{b}_1 x_{1i..}-\hat{b}_2 x_{2i..}$$

$$C_1\sigma_{y_{i..}y_{.i.}}+C_2\sigma^2_{y_{.i.}}=y_{.i.}-\hat{b}_1 x_{1.i.}-\hat{b}_2 x_{2.i.}$$

The variances and covariances needed in this approximate solution can be computed easily from σ_g^2, σ_m^2, σ_s^2, and σ_e^2. Approximate values of $\hat{s}_{ij}$ can then be obtained by substituting the approximate $\hat{b}_1, \hat{b}_2, \hat{g}_i$, and $\hat{m}_j$ in equations (9).

ESTIMATION OF VARIANCES OF GENERAL, MATERNAL, AND SPECIFIC EFFECTS

As mentioned earlier, one might take as the additive genetic variance and covariance among the lines the theoretical values based on relationships among the lines, degree of inbreeding among the lines, and the genetic variance in the original population from which the lines came. It is necessary even then to estimate σ_m^2, σ_s^2, and σ_e^2. It is well known that methods for estimating variance components are in a much less advanced stage than estimation of individual fixed effects. It is seldom possible to obtain maximum likelihood estimates. Consequently many different methods might be used, and the relative efficiencies of alternative procedures are not known.

We shall consider as desirable criteria of estimation procedures for variance components ease of computation and unbiasedness. If the single cross experiment is a balanced one, that is if there are the same number of observations on each of the possible crosses, it is not difficult to work out the least squares sums of squares for various tests of hypotheses, regarding the line and cross line characteristics as fixed. Then assuming that there are no covariances between the various effects and interactions, one can obtain the expectations of the least squares sum of squares under the assumption that the effects and interactions have a distribution (Henderson, 1948). In case the experiment is not a balanced one, it is still possible to obtain least squares tests of hypotheses and to find expectations of the resulting sums of squares. This, however, is ordinarily an extremely laborious procedure (Henderson, 1950).

A much easier procedure is available. It probably gives estimates with larger sampling variance, although that is not really known, and gives almost exactly the same results in the balanced experiments as does the least squares procedure. This involves computing various sums of squares ignoring all criteria of classification except one, taking expectations of these various sums of squares, and solving the resulting set of simultaneous equations. The latter procedure will now be illustrated for single cross data in which we wish to obtain estimates of the variances pertaining to general combining ability, maternal ability, specific effects, and error. It will be assumed that the only fixed element in the model is μ. Now let us compute certain sums of squares and their expectations. These are set out below.

$$\text{Total: } E\left(\sum_i \sum_j \sum_k y_{ijk}^2\right) = n_{\cdot\cdot}(\mu^2 + 2\sigma_g^2 + \sigma_m^2 + \sigma_s^2 + \sigma_e^2)$$

$$\text{Sires: } E\left(\sum_i \frac{y_{i..}^2}{n_{i.}}\right) = n..(\mu^2+\sigma_g^2) + \sum_i \frac{\sum_j n_{ij}^2}{n_{i.}}(\sigma_g^2+\sigma_m^2+\sigma_s^2) + s\sigma_e^2,$$

where s denotes number of different lines used as the male line.

$$\text{Dams: } E\left(\sum_j \frac{y_{.j.}^2}{n_{.j}}\right) = n..(\mu^2+\sigma_g^2+\sigma_m^2) + \sum_j \frac{\sum_i n_{ij}^2}{n_{.j}}(\sigma_g^2+\sigma_s^2) + d\sigma_e^2,$$

where d is the number of different lines used as the female line.

$$\text{Crosses: } E\left(\sum_{i<j} \frac{(y_{ij.}+y_{ji.})^2}{n_{ij}+n_{ji}}\right) = n..(\mu^2+2\sigma_g^2+\sigma_s^2) + \sum_{i<j} \frac{n_{ij}^2+n_{ji}^2}{n_{ij}+n_{ji}}\sigma_m^2 + c\sigma_e^2,$$

where c denotes the number of different crosses (regarding reciprocals as one cross)

$$\text{Correction Factor: } E\left(\frac{y_{...}^2}{n..}\right) = n..\mu^2 + \sum_i (n_{i.}+n_{.i})^2\sigma_g^2/n.. + \sum_j n_{.j}^2\sigma_m^2/n.. + \sum_{i<j} (n_{ij}+n_{ji})^2\sigma_s^2/n.. + \sigma_e^2$$

The above sums of squares and expectations are quite easy to compute and once this is done all one needs to do is to subtract the correction factor and its expectation from the other sums of squares and expectations and solve the resulting set of four equations for σ_g^2, σ_m^2, σ_s^2, and σ_e^2.

FURTHER RESEARCH NEEDED

If maximum progress through selection for general and specific combining ability is to be attained, much additional research is needed. From a statistical standpoint we need to know if an index based on minimization of $E(\hat{\theta}-\theta)^2$ comes close to maximizing progress through selection by truncation when the distributions are not the multivariate normal. If such an index does not do so, we need to know what practicable index or indexes will. Further, if nothing is known of the variances and covariances needed in construction of indexes or if there are available only estimates with large sampling errors, we need to know if the index based on the assumption that the estimate is the true value is best from the standpoint of maximizing genetic progress. Finally, much more work is needed on the problem of estimating variance and covariance components and placing confidence limits on such estimates.

Although there is a considerable body of literature on heritability estimates, we need more accurate estimates of the heritabilities of most traits of economic importance. Also almost nothing is now known about genetic correlations between traits, about genetic-environmental interactions, and about the magnitude of genetic differences among herds. Estimates of these genetic parameters are essential to intelligent selection for additive genetic values. In the case of inbred lines, little is known concerning the variances of general and specific combining abilities. The work of Sprague and Tatum (1942) with corn and Henderson (1949) with swine illustrates the types of estimates which are badly needed in selecting for general and specific combining abilities from the results of line cross tests.

Finally, well designed experiments are needed to test how closely predictions made from indexes or other selection procedures check with actual results.