Prediction of Double Crosses from Single Crosses*[1]

C. CLARK COCKERHAM

Department of Experimental Statistics, North Carolina State University at Raleigh

Summary. A unified theory, which takes into account both the genetic and experimental conditions, for predicting double cross hybrids from single cross hybrids is developed. The method is analogous to that for selection indices. Relationship of the prediction model to the genetic model is explored. JENKINS' (1934) three single cross predictors, a best single cross predictor, and selection on the basis of double cross estimates are compared empirically for an additive and dominance genetic model with varying proportions of experimental error variance and different numbers of hybrids. The differences between fixed and random genetic sample approaches to prediction are discussed.

With many inbred lines it is often impractical to test and compare all possible double cross hybrids. The number of single crosses is considerably fewer, however, and it is logical that they can be used to estimate or predict the performance of the double cross. Questions arise then as to the best methods of prediction and the accuracy of the predictions.

Table 1. *Analysis of variance of all possible single crosses.*

Source	Degrees of Freedom	Mean Square	Expectation of mean square
Replications	$r-1$		
General	$p-1$	$\hat{M}_3$	$M_3 = \sigma^2 + r\,\sigma_s^2 + r\,(p-2)\,\sigma_g^2$
Specific	$p\,(p-3)/2$	$\hat{M}_2$	$M_2 = \sigma^2 + r\,\sigma_s^2$
Error	$(r-1)\,(p-2)\,(p+1)/2$	$\hat{M}_1$	$M_1 = \sigma^2$

The logic of predicting double cross hybrids from single cross hybrids was recognized early in the development of double cross hybrids in corn. JENKINS (1934) presented four alternative methods of predicting double cross performance, three of which involved single crosses, and the first favorable evidence for their use. Since then several studies of one or more of the methods have been reported (reviewed by SPRAGUE, 1955).

EBERHART (1964) gave theoretical relations among hybrids and considered two of JENKINS' (1934) predictors in conjunction with information on three-way crosses in the prediction of double cross performance. EBERHART et al. (1964) compared predicted values of double crosses in maize for various predictors, but did not have double cross information to evaluate the methods. EBERHART's procedures were based on what shall be referred to as a fixed sample approach. In contrast, the theory in the present study, which also takes into account both the genetic and experimental conditions, is based on a random sample approach. The two approaches, both of which have been used extensively in quantitative genetics and breeding methodology, will be discussed more fully at the end.

Single Cross Experiment

To furnish a base for familiarity, consider an experiment of r replications of all possible single crosses from p inbred lines. This is the common replicated diallel experiment without reciprocals and parent lines, for which the analysis of variance of plot values is that given in Table 1. The components of variance in Table 1 are σ^2 for error, σ_s^2 for specific combining ability, and σ_g^2 for general combining ability.

Single Cross Information

For each possible double cross the single crosses may be grouped according to their relationship to the double cross. To the double cross $AB \cdot CD$, for example, the two parental single crosses, AB and CD, bear the same relationship; the four nonparental single crosses, AC, AD, BC and BD, bear the same relationship; the single crosses, AE, AF, ..., BE, BF, ..., CE, CF, ..., DE, DF, ..., involving only one of the lines, A, B, C and D, bear the same relationship; and the single crosses involving none of the four parent lines bear the same relationship, being zero in this latter case. Consequently, for each double cross, the following four means of the single crosses are considered:

X_p = the mean of the two parental single crosses; *e.g.*, $(\overline{AB} + \overline{CD})/2$.

X_n = the mean of the four nonparental single crosses; *e.g.*, $(\overline{AC} + \overline{AD} + \overline{BC} + \overline{BD})/4$.

X_l = the mean of the single crosses from the four parental lines in combination with all other lines; *e.g.*, $(\overline{AE} + \overline{AF} + \cdots + \overline{BE} + \overline{BF} + \cdots + \overline{CE} + \overline{CF} + \cdots + \overline{DE} + \overline{DF} + \cdots)/4\,(p-4)$. (1)

X_0 = the mean of the single crosses from the other lines in all possible combinations; *e.g.*, $(\overline{EF} + \overline{EG} + \overline{EH} + \cdots + \overline{FG} + \overline{FH} + \cdots + \overline{GH} + \cdots)/(p-4)\,(p-5)/2$.

The divisors indicate the number of single crosses in each mean, and the bar over each single cross indicates that it is a mean for the experiment (r replications in the example). These means are dependent in that

$$\overline{X} = [2\,X_p + 4\,X_n + 4\,(p-4)\,X_l + (p-4)\,(p-5)\,X_0/2]/p\,(p-1)/2 \quad (2)$$

* Dedicated to Dr. GEORGE F. SPRAGUE on the occasion of his 65th birthday.

[1] Contribution from the Department of Experimental Statistics, N. C. Agricultural Experiment Station, Raleigh, North Carolina. Published with the approval of the Director of Research as Paper No. 2330 of the Journal Series. This investigation was supported in part by Public Health Service Research Grant GM 11546 from the Division of General Medical Sciences.

is the mean of all the single crosses and is the same for each double cross. Thus, only three not wholly dependent kinds of single cross information are available for a given double cross.

It is helpful to define two additional means of the single crosses for each double cross.

$$\left.\begin{aligned} X_{\bar{p}} &= (X_p + 2X_n)/3 = \text{the mean of the six single crosses involving the four parental lines of the double cross.} \\ X_{\bar{l}} &= [X_p + 2X_n + (p-4)X_l]/(p-1) = \text{the mean of the four parental line means where a parental line mean is the mean of all } (p-1) \text{ single crosses involving that line.} \end{aligned}\right\} \quad (3)$$

Linear Prediction Equation

The result desired is the linear combination of the three kinds of information, Z's, on the single crosses which best predicts the performances, Y's, of the double crosses.

$$Y = \mu + \hat{Y} + e = \mu + b_1 Z_1 + b_2 Z_2 + b_3 Z_3 + e\,. \quad (4)$$

The b's are the relative weights to be given to the single cross information and $e = Y - \mu - \hat{Y}$ is the error of prediction. The mean of the double crosses, μ, plays no part in the relative ranking of the double crosses.

Within limits, the choice of the Z's is a matter of convenience. In the following form they are uncorrelated,

$$\left.\begin{aligned} \underline{Z}' = (Z_1 &= X_n - X_{\bar{p}}\,, \\ Z_2 &= X_{\bar{p}} - 2X_l + X_0\,, \\ Z_3 &= X_{\bar{l}} - \overline{X})\,, \end{aligned}\right\} \quad (5)$$

and have variances, $\sigma^2_{Z_i} = E(Z_i^2)$,

$$\left.\begin{aligned} \sigma^2_{Z_1} &= \left(\frac{\sigma^2}{r} + \sigma_s^2\right)\frac{1}{12} = \frac{M_2}{r}\frac{1}{12}\,, \\ \sigma^2_{Z_2} &= \left(\frac{\sigma^2}{r} + \sigma_s^2\right)\frac{(p-1)(p-2)}{6(p-4)(p-5)} \\ &= \frac{M_2}{r}\frac{(p-1)(p-2)}{6(p-4)(p-5)}\,, \\ \sigma^2_{Z_3} &= \left[\frac{\sigma^2}{r} + \sigma_s^2 + (p-2)\sigma_g^2\right]\frac{(p-2)(p-4)}{4p(p-1)^2} \\ &= \frac{M_3}{r}\frac{(p-2)(p-4)}{4p(p-1)^2}\,. \end{aligned}\right\} \quad (6)$$

In matrix notation, the variance covariance matrix is

$$\sigma_{ZZ} = \begin{bmatrix} \sigma^2_{Z_1} & 0 & 0 \\ 0 & \sigma^2_{Z_2} & 0 \\ 0 & 0 & \sigma^2_{Z_3} \end{bmatrix}. \quad (7)$$

Let us write formally, for the moment, the covariances of the information on the single crosses with the double crosses,

$$\underline{\sigma}'_{ZY} = (\sigma_{Z_1 Y}, \sigma_{Z_2 Y}, \sigma_{Z_3 Y})\,. \quad (8)$$

To be used as a criterion of the best predictor is that set of b's which minimizes the variance of the prediction error, $\sigma_e^2 = E(Y - \mu - \hat{Y})^2$. The minimization procedure is straightforward, analogous to least squares. The b's,

$$\underline{b}^{*\prime} = (b_1^*, b_2^*, b_3^*)\,, \quad (9)$$

that satisfy this criterion are solutions to the following set of equations,

$$\sigma_{ZZ}\,\underline{b}^* = \underline{\sigma}_{ZY}\,, \quad (10)$$

and

$$b_1^* = \frac{\sigma_{Z_1 Y}}{\sigma^2_{Z_1}}\,, \quad b_2^* = \frac{\sigma_{Z_2 Y}}{\sigma^2_{Z_2}}\,, \quad b_3^* = \frac{\sigma_{Z_3 Y}}{\sigma^2_{Z_3}}\,. \quad (11)$$

For any given predictor, $\hat{Y}_i = \underline{b}^{i\prime}\underline{Z}$, let $\overline{Y}_{is}$ be the mean of the set of double crosses corresponding to a selected set of $\hat{Y}_i$ with mean $\overline{\hat{Y}}_{is}$, and let $\overline{Y}$ and $\overline{\hat{Y}}_i$ be the corresponding means for the entire sample. Based on linear regression, which to be exact requires normal distribution of X and Y,

$$\left.\begin{aligned} \Delta_i &= \overline{Y}_{is} - \overline{Y} = B_{Y\hat{Y}_i}(\overline{\hat{Y}}_{is} - \overline{\hat{Y}}_i) \\ B_{Y\hat{Y}_i} &= \frac{\operatorname{Cov} Y\hat{Y}_i}{\sigma^2_{\hat{Y}_i}} = \frac{\underline{b}^{i\prime}\underline{\sigma}_{ZY}}{\underline{b}^{i\prime}\sigma_{ZZ}\underline{b}^i}\,, \end{aligned}\right\} \quad (12)$$

where $B_{Y\hat{Y}_i}$ is the regression coefficient of Y on $\hat{Y}_i$. For truncation selection one may substitute

$$k\,\sigma_{\hat{Y}_i} = k\sqrt{\underline{b}^{i\prime}\sigma_{ZZ}\underline{b}^i} = \overline{\hat{Y}}_{is} - \overline{\hat{Y}}_i \quad (13)$$

into (12), and

$$\Delta_i = k\frac{\underline{b}^{i\prime}\underline{\sigma}_{ZY}}{\sqrt{\underline{b}^{i\prime}\sigma_{ZZ}\underline{b}^i}} \quad (14)$$

where k is a function of the intensity of selection. The above relationships (12, 14) hold for any set of b's. For the set, $\underline{b}^*$, which minimizes the prediction error, substitution of (10) into (14) leads to

$$\Delta_* = k\sqrt{\underline{b}^{*\prime}\sigma_{ZZ}\underline{b}^*}\,, \quad (15)$$

and Δ is a maximum.

Genetic Interpretation

It will be mentioned now, and discussed later, that inherent in the previous definition and solution to the best predictor is the assumption that the single crosses and double crosses are random samples for which the procedure will be used. Under certain assumptions for diploids (COCKERHAM, 1963), mainly random samples from a noninbred population and no linkages, the covariances among relatives can be expressed in the following form,

$$C = \alpha\,\sigma_A^2 + \delta\,\sigma_D^2 + \alpha^2\sigma^2_{AA} + \alpha\,\delta\,\sigma^2_{AD} + \ldots\,, \quad (16)$$

where the subscripts, A = additive, D = dominance, indicate the various components of genetic variance. The coefficients α and δ are given by COCKERHAM

Covariance	α	δ	Description
C_{S_2}	1	1	— Covariance between single cross relatives with both parents common.
C_{S_1}	1/2	0	— Covariance between single cross relatives with one parent common only.
C_{SD_2}	1/2	1/4	— Covariance between nonparental single cross and double cross relatives, *i. e.*, AB with $A\text{-}\cdot B\text{-}$.
C_{SD_2}	1/2	0	— Covariance between parental single cross and double cross relatives, *i. e.*, AB with $AB\cdot\text{--}$. Note: $C_{SD_2} = C_{S_1}$.
C_{SD_1}	1/4	0	— Covariance between single cross and double cross relatives with one parent line common only.

(17)

(1961) for all possible types of single, three-way and double cross relatives. Although a more general solution for the coefficients with partially inbred parental lines has been found (COCKERHAM, 1967), we shall consider the parental lines to be homozygous, and shall reproduce here the coefficients for the hybrid relatives under consideration.

It remains to relate the covariances to the statistical parameters. These are

$$\left.\begin{aligned} \sigma_g^2 &= C_{S1}\,, \\ \sigma_s^2 &= C_{S2} - 2\,C_{S1}\,, \\ \sigma_{Z_1 Y} &= (C_{SD\bar{2}} - C_{SD2})/3\,, \\ \sigma_{Z_2 Y} &= 2\,\sigma_{Z_1 Y} + (C_{SD2} - 2\,C_{SD1})\,, \\ \sigma_{Z_3 Y} &= [3\,\sigma_{Z_2 Y} + (p-2)\,C_{SD1}]\,\frac{p-4}{p\,(p-1)}\,. \end{aligned}\right\} \quad (18)$$

Substitution of the appropriate coefficients, α and δ, in (17) into (16) gives the genetic variance composition of the covariances. Further substitution into (18) and then into (11) leads to expressions for the b's in terms of genetic and environmental variances and the experimental dimensions, p and r. While these expressions are not simple, interrelationships of the genetic and prediction models are illuminating.

The following alternative expressions of the b's will be helpful in genetic clarification,

$$\left.\begin{aligned} b_1^* &= \frac{4\,(C_{SD\bar{2}} - C_{SD2})}{M_2/r} = \frac{H_1}{M_2/r}\,, \\ b_2^* &= \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)} \times \\ &\quad \times \left[\frac{4\,(C_{SD\bar{2}} - C_{SD2}) + 6\,(C_{SD2} - 2\,C_{SD1})}{M_2/r}\right] \\ &= \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}\left[b_1 + \frac{6\,(C_{SD2} - 2\,C_{SD1})}{M_2/r}\right] \\ &= \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}\,\frac{H_2}{M_2/r}\,, \\ b_3^* &= \frac{\left\{\begin{array}{c}(p-1)\,[8\,(C_{SD\bar{2}} - C_{SD2}) + 12\,(C_{SD2} - 2\,C_{SD1})] \\ + 4\,(p-2)\,C_{SD1}\end{array}\right\}}{(p-2)\,M_3/r} \\ &= \frac{(p-1)^2\,2\,b_2\,M_2}{(p-4)\,(p-5)\,M_3} + \frac{4\,(p-1)\,C_{SD1}}{M_3/r} \\ &= \frac{(p-1)\,H_3}{(p-2)\,(M_3/r)}\,. \end{aligned}\right\} \quad (20)$$

The genetic variance compositions of the various covariances of relatives, and functions of them, are listed in Table 2 for ready reference. A glance at the genetic numerators, H's, of the b's shows that for b_1^* and b_2^* to be other than 0 requires some dominance and/or epistasis. Additive types of epistasis contribute to H_2 but not to H_1. The dominating coefficients in H_3 are for the all-additive types of variance for large p but all variances are included to some degree.

Table 2. *Genetic variance composition of various functions of the covariances of relatives.*

Covariance	Genetic Components of Variance									
	σ_A^2	σ_D^2	σ_{AA}^2	σ_{AD}^2	σ_{DD}^2	σ_{AAA}^2	σ_{AAD}^2	σ_{ADD}^2	σ_{DDD}^2	...
$C_{S1} = C_{SD2}$	$\frac{1}{2}$	0	$\frac{1}{4}$	0	0	$\frac{1}{8}$	0	0	0	...
C_{S2}	1	1	1	1	1	1	1	1	1	...
C_{SD1}	$\frac{1}{4}$	0	$\frac{1}{16}$	0	0	$\frac{1}{64}$	0	0	0	...
$C_{SD\bar{2}}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	...
$C_{S2} - 2\,C_{S1}$	0	1	$\frac{1}{2}$	1	1	$\frac{3}{4}$	1	1	1	...
$C_{SD2} - 2\,C_{SD1}$	0	0	$\frac{1}{8}$	0	0	$\frac{3}{32}$	0	0	0	...
$C_{SD\bar{2}} - C_{SD2}$	0	$\frac{1}{4}$	0	$\frac{1}{8}$	$\frac{1}{16}$	0	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{64}$	...
H_1	0	1	0	$\frac{1}{2}$	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	...
H_2	0	1	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{9}{16}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	...
H_3	$(p-2)$	2	$\frac{p+4}{4}$	1	$\frac{1}{2}$	$\frac{p+16}{16}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	...

Relationship of the Prediction to the Genetic Model

First of all, note that Z_2 and Z_3 are constant for the three orders, $AB \cdot CD$, $AC \cdot BD$ and $AD \cdot BC$, of double crosses for any four lines. Consequently, the only information about these comparisons is contained in Z_1 and the relative importance of this information is b_1. The information in Z_2 is of an interaction form, involving the mean of all combinations of four lines, $X_{\bar{p}}$, the mean of the four lines in combination with all other lines, X_l, and the mean of all other lines in combination, X_0. A two by two table illustrates the point.

	4 lines	Other lines
4 lines	$X_{\bar{p}}$	X_l
Other lines	X_l	X_0

(19)

It, $Z_2 = X_{\bar{p}} - 2\,X_l + X_0$, is a measure of the interaction of the four lines with the other lines, and the importance of this general nicking of four lines in predicting the corresponding double cross performance is reflected by b_2. The relative importance of Z_3 depends on how well a line's performance in all single crosses relates to the double cross performance.

Since none of the differences among the covariances in (20) can be negative, the following relationships must hold,

$$\left.\begin{aligned} &0 \le b_1^* \le \frac{1}{2}\,, \quad b_1^* \le \frac{(p-1)\,(p-2)}{(p-4)\,(p-5)}\,b_2^* \le \frac{3}{2}\,, \\ &b_2^*\,\frac{12\,(p-1)\,M_2}{(p-2)\,M_3} \le b_3^* \le \frac{2\,(p-1)}{p-2}\,. \end{aligned}\right\} \quad (21)$$

While these relationships do not definitely establish an order among the b's, they have implications for practical applications.

Simplification in the genetic model leads to simplification in the prediction model. If all of the genetic variance is additive variance, the b's reduce to

$$\underline{b}^1 = \left[0, 0, \frac{(p-1)\,\sigma_A^2}{(\sigma^2/r) + (p-2)\,\sigma_A^2}\right]. \tag{22}$$

Since only proportionality of the b's are important,

$$\underline{b}^1 \propto (0, 0, 1)\,, \tag{23}$$

and the best prediction model involves only the parental line means,

$$\hat{Y}_1 = X_{\bar{i}} - \bar{X}\,, \tag{24}$$

which is one of the predictors proposed by JENKINS (1934) (his method C). Thus, this predictor is best only when there are only additive effects of genes.

Another of JENKINS' predictors (his method B) is best when all of the genetic variance is dominance variance,

$$\underline{b}^2 = \left[\frac{\sigma_D^2}{(\sigma^2/r) + \sigma_D^2}, \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}\,\frac{\sigma_D^2}{((\sigma^2/r) + \sigma_D^2)}, \frac{2\,(p-1)}{(p-2)}\,\frac{\sigma_D^2}{((\sigma^2/r) + \sigma_D^2)}\right], \tag{25}$$

or

$$\underline{b}^2 \propto \left[1, \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}, \frac{2\,(p-1)}{p-2}\right], \tag{26}$$

and the prediction equation is

$$\hat{Y}_2 = X_n - \bar{\bar{X}}\,, \tag{27}$$

which involves only the nonparental single crosses. A comment is in order. The definitions of genetic effects and variances are such that for there to be only dominance variance requires overdominance at all loci, and further, the noninbred population must have equilibrium frequencies at each locus, a very unlikely condition. There is another case, however, of more importance for which $\underline{b}^2$ as in (26) is equivalent to $\underline{b}^*$. This is when all of the genetic variance is additive and dominance and the experimental error variance is zero, *i. e.*, $\sigma^2/r = 0$. Thus, with increasing r, $\underline{b}^2$ tends toward the best predictor when there are both additive and dominance effects of genes but no epistatic effects.

JENKINS' third method (A) of utilizing single crosses assumes the following set of b's,

$$\underline{b}^3 \propto \left(0, \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}, \frac{2\,(p-1)}{p-2}\right), \tag{28}$$

which corresponds to the following prediction equation,

$$\hat{Y}_3 = X_{\bar{p}} - \bar{X} \tag{29}$$

and amounts to using the mean of the six single crosses of the four parents of the double cross. There appears to be no simple genetic situation for which this predictor would be best.

If there are no epistatic effects, but allowing for both additive and dominance variance,

$$\underline{b}^4 = \left\{\frac{\sigma_D^2}{(\sigma^2/r) + \sigma_D^2}, \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}\,\frac{\sigma_D^2}{((\sigma^2/r) + \sigma_D^2)}, \frac{(p-1)\,[2\,\sigma_D^2 + (p-2)\,\sigma_A^2]}{(p-2)\,[(\sigma^2/r) + \sigma_D^2 + (p-2)\,\sigma_A^2/2]}\right\}, \tag{30}$$

or

$$\underline{b}^4 \propto \left\{1, \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}, \frac{(p-1)\,((\sigma^2/r) + \sigma_D^2)\,[2\,\sigma_D^2 + (p-2)\,\sigma_A^2]}{(p-2)\,\sigma^2{}_I\,[(\sigma^2/r) + \sigma_D^2 + (p-2)\,\sigma_A^2/2]}\right\}. \tag{31}$$

All of the parameters are contained in the diallel experiment. Consequently, the b's may be written as functions of the expected mean squares,

$$\underline{b}^4 = \left[\frac{M_2 - M_1}{M_2}, \frac{(p-4)\,(p-5)\,(M_2 - M_1)}{(p-1)\,(p-2)\,M_2}, \frac{(p-1)\,(M_3 + M_2 - 2\,M_1)}{(p-2)\,M_3}\right], \tag{32}$$

or

$$\underline{b}^4 \propto \left[1, \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}, \frac{2\,(p-1)\,M_2}{(p-2)\,M_3}\left(1 + \frac{M_3 - M_2}{M_2 - M_1}\right)\right]. \tag{33}$$

Thus, from a diallel experiment, assuming no epistasis, the b's may be estimated. The proportional b's in (31, 33) assume at least some dominance variance. Otherwise we have divided through by zero in standardizing b_1 to be one.

To admit all types of epistatic variance requires estimates not available in the diallel experiment, namely, H_1, H_2 and H_3 or two relative values, say H_2/H_1 and H_3/H_1. If there are no additive types of epistatic variance then $H_2 = H_1$ and $H_3 = 2\,H_1 + 2\,(M_3 - M_2)/r$, but still required is the relative value of $\sigma_A^2 = 2\,(M_3 - M_2)/(p-2)\,r$ to H_1. The proportional b's would be the same as in (33) except for b_3,

$$\underline{b}^5 \propto \left[1, \frac{(p-4)\,(p-5)}{(p-1)\,(p-2)}, \frac{2\,(p-1)\,M_2}{(p-2)\,M_3}\left(1 + \frac{M_3 - M_2}{r\,H_1}\right)\right], \tag{34}$$

and b_3 would be relatively larger when there was epistasis involving dominance instead of only dominance effects.

Goodness of Predictors

The appeal of best may not always be warranted. Sometimes a best predictor as defined herein is only infinitesimally better than another one, and all may be good or poor depending on the parametric situation.

We shall compare predictors empirically as ratios of their expected gains, Δ's, to gain expected from selecting among the true values, Y's, of the double crosses. The gain expected from selecting among the Y's,

$$\Delta_Y = k\,\sigma_Y\,, \tag{35}$$

is dependent upon the selection intensity reflected in k and σ_Y^2 which is derived in an appendix. The efficiency measure,

$$\frac{\Delta_i}{\Delta_Y} = \frac{\underline{b}^{i\prime}\,\sigma_{ZY}}{\sqrt{\underline{b}^{i\prime}\,\sigma_{ZZ}\,\underline{b}^i}\,\sigma_Y} = \varrho_{iY}\,, \tag{36}$$

is also the correlation, ϱ_{iY}, between the predictor, $\hat{Y}_i$, and Y, which can be verified from relationships given in (12). The relative efficiency of any two predictors, $\hat{Y}_i$ and $\hat{Y}_j$, may be obtained simply as

$$\frac{\Delta_i}{\Delta_j} = \frac{\varrho_{iY}}{\varrho_{jY}} \tag{37}$$

which, when $j = *$,

$$\frac{\Delta_i}{\Delta_*} = \frac{\varrho_{iY}}{\varrho_{*Y}} = \varrho_{i*}\,, \tag{38}$$

is also the correlation between $\hat{Y}_i$ and $\hat{Y}_*$.

Since the true genotypic values of the double crosses are never known, we wish to know also how

effective is selection based on experimental estimates of double cross performance as a further basis of comparison. Among many alternatives we shall consider that the same amount of experimental information is available for the double crosses as for the single crosses, *i. e.*, the same number, r, of replicate plots for each double cross. Denoting the mean over replications of individual double crosses as $\tilde{Y}$,

$$\tilde{Y} = Y + \bar{\varepsilon}, \tag{39}$$

where $\bar{\varepsilon}$ is the experimental error. The gain from selecting on $\tilde{Y}$, following arguments (12) through (14), is

$$\Delta_{\tilde{Y}} = k\frac{\sigma^2_{\tilde{Y}}}{\sigma_{\tilde{Y}}}. \tag{40}$$

The variance, $\sigma^2_{\tilde{Y}}$, among the estimates is derived simultaneously with σ^2_Y in the Appendix.

The ratio of the gain (40) from using the estimates to that (35) from knowing the true values,

$$\frac{\Delta_{\tilde{Y}}}{\Delta_Y} = \frac{\sigma_Y}{\sigma_{\tilde{Y}}} = \varrho_{\tilde{Y}Y}, \tag{41}$$

is again the correlation between the two, which approaches one as the experimental error variance approaches zero.

To be compared empirically are JENKINS' (1934) three predictors (23, 26 and 28) and the best predictor, $\hat{Y}_*$. Their efficiencies in the form of (36) and the efficiency of experimental double cross prediction (41) give five comparative measures,

$$\varrho_{1Y},\ \varrho_{2Y},\ \varrho_{3Y},\ \varrho_{*Y},\ \varrho_{\tilde{Y}Y}. \tag{42}$$

These efficiency measures are affected only by changes in the relative values of the variance parameters. Consequently, if there are no epistatic effects, the variances can be reduced to two quantities,

$$\beta = \frac{\sigma^2_A + \sigma^2_D}{(\sigma^2/r) + \sigma^2_A + \sigma^2_D}, \quad \gamma = \frac{\sigma^2_A}{\sigma^2_A + \sigma^2_D}, \tag{43}$$

such that

$$\left[\frac{\sigma^2}{r} = 1 - \beta\right] + [\sigma^2_D = \beta\,(1-\gamma)] + [\sigma^2_A = \gamma\,\beta] = 1. \tag{44}$$

The proportion of the total variance that is genotypic is measured by β, and of the genotypic variance that is additive by γ.

The five efficiency measures (42) are given for various combinations of γ and β in Table 3 for $p = 6$ and $p = 20$, and in Table 4 for $p = 100$ and $p = \infty$. Also included in the Tables is σ_Y to show how it, and consequently Δ_Y in (35), varies with the parameters. While the error variance, σ^2_ε, in $\sigma^2_{\tilde{Y}}$ for double crosses is slightly larger than σ^2 for the single crosses, (74) and (75), the two were taken to be the same which has a very slight inflationary effect on $\varrho_{\tilde{Y}Y}$.

First, consider the ϱ_{iY}'s in Tables 3 and 4 relating to single cross prediction only. As a base of reference, $\hat{Y}_*$ is an ideal and represents the best that one can do with single cross prediction. Its efficiency, ϱ_{*Y}, increases with γ, with β, and with p except for $\gamma = 0$. There is an overriding effect of β, however, in that variations in γ have less of an effect as β

Table 3. *Efficiencies of Predictors for $p = 6$ and $p = 20$ (in parentheses).*

β		γ = 0.0	0.1	0.3	0.5	0.7	0.9	1.0
0.1	ϱ_{1Y}	.21 (.13)	.23 (.23)	.28 (.38)	.32 (.50)	.36 (.59)	.40 (.67)	.43 (.71)
	ϱ_{2Y}	.32 (.32)	.31 (.33)	.31 (.34)	.31 (.36)	.31 (.37)	.30 (.39)	.30 (.39)
	ϱ_{3Y}	.23 (.26)	.25 (.28)	.28 (.33)	.31 (.38)	.35 (.41)	.38 (.45)	.40 (.46)
	ϱ_{*Y}	.32 (.32)	.32 (.34)	.32 (.42)	.34 (.51)	.37 (.60)	.41 (.67)	.43 (.71)
	$\varrho_{\tilde{Y}Y}$	.14 (.16)	.14 (.17)	.14 (.18)	.14 (.19)	.14 (.19)	.14 (.20)	.14 (.21)
	σ_Y	.14 (.16)	.13 (.16)	.13 (.17)	.13 (.18)	.13 (.19)	.13 (.20)	.13 (.20)
0.3	ϱ_{1Y}	.37 (.23)	.40 (.37)	.47 (.55)	.53 (.69)	.59 (.78)	.65 (.86)	.68 (.89)
	ϱ_{2Y}	.55 (.55)	.55 (.56)	.54 (.58)	.54 (.60)	.54 (.62)	.53 (.63)	.53 (.64)
	ϱ_{3Y}	.40 (.44)	.43 (.48)	.48 (.55)	.53 (.61)	.57 (.66)	.62 (.70)	.65 (.72)
	ϱ_{*Y}	.55 (.55)	.55 (.57)	.55 (.65)	.57 (.72)	.61 (.80)	.65 (.86)	.68 (.89)
	$\varrho_{\tilde{Y}Y}$	.27 (.31)	.27 (.32)	.27 (.33)	.27 (.35)	.26 (.36)	.26 (.38)	.26 (.38)
	σ_Y	.23 (.27)	.23 (.28)	.23 (.30)	.23 (.31)	.23 (.33)	.22 (.34)	.22 (.35)
0.5	ϱ_{1Y}	.48 (.30)	.51 (.44)	.59 (.63)	.65 (.76)	.72 (.85)	.78 (.92)	.82 (.95)
	ϱ_{2Y}	.71 (.71)	.71 (.72)	.70 (.74)	.70 (.75)	.70 (.77)	.69 (.78)	.69 (.79)
	ϱ_{3Y}	.52 (.57)	.55 (.62)	.61 (.69)	.66 (.74)	.71 (.79)	.76 (.83)	.79 (.84)
	ϱ_{*Y}	.71 (.71)	.71 (.73)	.71 (.77)	.72 (.82)	.75 (.87)	.79 (.92)	.82 (.95)
	$\varrho_{\tilde{Y}Y}$	.40 (.44)	.40 (.45)	.39 (.48)	.39 (.49)	.39 (.51)	.38 (.53)	.38 (.53)
	σ_Y	.30 (.35)	.30 (.36)	.30 (.38)	.30 (.40)	.29 (.42)	.29 (.44)	.29 (.45)
0.7	ϱ_{1Y}	.56 (.36)	.60 (.50)	.68 (.68)	.75 (.80)	.81 (.88)	.88 (.95)	.91 (.98)
	ϱ_{2Y}	.84 (.84)	.84 (.84)	.83 (.86)	.83 (.87)	.83 (.88)	.83 (.89)	.82 (.89)
	ϱ_{3Y}	.62 (.68)	.65 (.72)	.70 (.79)	.76 (.84)	.81 (.88)	.87 (.91)	.89 (.92)
	ϱ_{*Y}	.84 (.84)	.84 (.85)	.84 (.87)	.84 (.90)	.86 (.92)	.88 (.95)	.91 (.98)
	$\varrho_{\tilde{Y}Y}$	.55 (.60)	.55 (.61)	.55 (.64)	.54 (.66)	.54 (.67)	.53 (.69)	.53 (.69)
	$\sigma_{\tilde{Y}}$	.36 (.41)	.36 (.43)	.35 (.45)	.35 (.48)	.35 (.50)	.34 (.52)	.34 (.53)
0.9	ϱ_{1Y}	.64 (.40)	.68 (.54)	.75 (.71)	.82 (.82)	.88 (.90)	.94 (.97)	.97 (.99)
	ϱ_{2Y}	.95 (.95)	.95 (.95)	.95 (.96)	.95 (.96)	.95 (.96)	.94 (.97)	.94 (.97)
	ϱ_{3Y}	.70 (.77)	.73 (.81)	.78 (.86)	.84 (.90)	.89 (.94)	.94 (.97)	.97 (.98)
	ϱ_{*Y}	.95 (.95)	.95 (.95)	.95 (.96)	.95 (.96)	.95 (.97)	.96 (.98)	.97 (.99)
	$\varrho_{\tilde{Y}Y}$	.79 (.83)	.79 (.84)	.79 (.85)	.79 (.86)	.78 (.87)	.78 (.88)	.78 (.88)
	σ_Y	.41 (.47)	.40 (.48)	.40 (.51)	.40 (.54)	.39 (.56)	.39 (.59)	.39 (.60)
1.0	ϱ_{1Y}	.67 (.43)	.71 (.55)	.78 (.72)	.85 (.83)	.91 (.91)	.97 (.97)	1.00 (1.00)
	ϱ_{2Y}	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	ϱ_{3Y}	.74 (.81)	.77 (.84)	.82 (.89)	.87 (.93)	.92 (.96)	.97 (.99)	1.00 (1.00)
	ϱ_{*Y}	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	$\varrho_{\tilde{Y}Y}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	σ_Y	.43 (.49)	.43 (.51)	.42 (.54)	.42 (.57)	.41 (.59)	.41 (.62)	.41 (.63)

Table 4. *Efficiencies of Predictors for* $p = 100$ *and* $p = \infty$ (*in parentheses*).

		γ 0.0	0.1	0.3	0.5	0.7	0.9	1.0
0.1	ϱ_{1Y}	.06 (.00)	.29 (.43)	.55 (.68)	.70 (.82)	.81 (.91)	.89 (.97)	.92 (1.00)
	ϱ_{2Y}	.32 (.32)	.33 (.33)	.35 (.36)	.37 (.38)	.39 (.40)	.41 (.42)	.42 (.43)
	ϱ_{3Y}	.26 (.26)	.29 (.29)	.35 (.35)	.40 (.40)	.44 (.45)	.48 (.48)	.49 (.50)
	ϱ_{*Y}	.32 (.32)	.39 (.51)	.58 (.71)	.71 (.83)	.81 (.91)	.89 (.97)	.92 (1.00)
	$\varrho_{\tilde{Y}Y}$	.16 (.16)	.17 (.17)	.19 (.19)	.20 (.20)	.21 (.21)	.22 (.22)	.23 (.23)
	σ_Y	.16 (.16)	.17 (.17)	.18 (.18)	.19 (.19)	.20 (.21)	.21 (.22)	.22 (.22)
0.3	ϱ_{1Y}	.11 (.00)	.39 (.43)	.64 (.68)	.78 (.82)	.88 (.91)	.95 (.97)	.98 (1.00)
	ϱ_{2Y}	.55 (.55)	.56 (.57)	.59 (.60)	.62 (.63)	.64 (.65)	.66 (.67)	.67 (.68)
	ϱ_{3Y}	.45 (.45)	.49 (.50)	.57 (.58)	.63 (.64)	.68 (.69)	.73 (.73)	.74 (.75)
	ϱ_{*Y}	.55 (.55)	.61 (.64)	.73 (.77)	.82 (.85)	.89 (.92)	.95 (.97)	.98 (1.00)
	$\varrho_{\tilde{Y}Y}$	.31 (.31)	.32 (.32)	.35 (.35)	.37 (.37)	.39 (.39)	.40 (.41)	.41 (.42)
	σ_Y	.27 (.27)	.29 (.29)	.31 (.31)	.33 (.34)	.35 (.36)	.37 (.38)	.38 (.39)
0.5	ϱ_{1Y}	.14 (.00)	.42 (.43)	.67 (.68)	.80 (.82)	.90 (.91)	.96 (.97)	.99 (1.00)
	ϱ_{2Y}	.71 (.71)	.72 (.72)	.75 (.75)	.77 (.77)	.79 (.79)	.80 (.81)	.81 (.82)
	ϱ_{3Y}	.58 (.58)	.63 (.63)	.71 (.71)	.76 (.77)	.81 (.82)	.85 (.85)	.86 (.87)
	ϱ_{*Y}	.71 (.71)	.74 (.75)	.81 (.83)	.87 (.88)	.92 (.93)	.97 (.98)	.99 (1.00)
	$\varrho_{\tilde{Y}Y}$	.45 (.45)	.46 (.46)	.49 (.50)	.52 (.52)	.54 (.55)	.56 (.57)	.57 (.58)
β	σ_Y	.35 (.35)	.37 (.37)	.40 (.40)	.43 (.43)	.45 (.46)	.48 (.49)	.49 (.50)
0.7	ϱ_{1Y}	.17 (.00)	.44 (.43)	.68 (.68)	.81 (.82)	.90 (.91)	.97 (.97)	1.00 (1.00)
	ϱ_{2Y}	.84 (.84)	.85 (.85)	.87 (.87)	.88 (.88)	.89 (.89)	.90 (.90)	.90 (.91)
	ϱ_{3Y}	.68 (.68)	.73 (.73)	.80 (.80)	.85 (.86)	.89 (.89)	.92 (.92)	.93 (.94)
	ϱ_{*Y}	.84 (.84)	.85 (.86)	.89 (.89)	.91 (.92)	.94 (.95)	.97 (.98)	1.00 (1.00)
	$\varrho_{\tilde{Y}Y}$	.61 (.61)	.62 (.63)	.65 (.66)	.68 (.68)	.70 (.71)	.72 (.73)	.73 (.73)
	σ_Y	.42 (.42)	.44 (.44)	.47 (.48)	.51 (.51)	.54 (.55)	.57 (.58)	.58 (.59)
0.9	ϱ_{1Y}	.19 (.00)	.45 (.43)	.69 (.68)	.82 (.82)	.91 (.91)	.97 (.97)	1.00 (1.00)
	ϱ_{2Y}	.95 (.95)	.95 (.95)	.96 (.96)	.96 (.96)	.97 (.97)	.97 (.97)	.97 (.97)
	ϱ_{3Y}	.77 (.77)	.81 (.82)	.87 (.88)	.92 (.92)	.95 (.95)	.97 (.97)	.98 (.98)
	ϱ_{*Y}	.95 (.95)	.95 (.95)	.96 (.96)	.97 (.97)	.97 (.98)	.98 (.99)	1.00 (1.00)
	$\varrho_{\tilde{Y}Y}$	.83 (.83)	.84 (.84)	.86 (.86)	.88 (.88)	.89 (.89)	.90 (.90)	.90 (.90)
	σ_Y	.47 (.47)	.50 (.50)	.54 (.54)	.57 (.58)	.61 (.62)	.64 (.65)	.66 (.67)
1.0	ϱ_{1Y}	.20 (.00)	.46 (.43)	.69 (.68)	.82 (.82)	.91 (.91)	.97 (.97)	1.00 (1.00)
	ϱ_{2Y}	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	ϱ_{3Y}	.82 (.82)	.85 (.85)	.90 (.91)	.94 (.94)	.97 (.97)	.99 (.99)	1.00 (1.00)
	ϱ_{*Y}	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	$\varrho_{\tilde{Y}Y}$	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)
	σ_Y	.50 (.50)	.52 (.52)	.56 (.57)	.60 (.61)	.64 (.65)	.68 (.69)	.69 (.71)

becomes larger. Turning now to $\hat{Y}_2$ we knew from (26), and the related comments, that it was equivalent to the most efficient predictor, $\hat{Y}_*$, when all of the genetic variance was dominance, *i. e.*, $\gamma = 0$, and when $\beta = 1$, *i. e.*, $\sigma^2/r = 0$. Actually, the efficiency is relatively insensitive to changes in γ and p, and mainly dependent on β. Varying p has no effect on the efficiency when $\gamma = 0$, and there is apparently a p between 6 and 20 for which variations in γ have an imperceptible, if any, effect on the efficiency because ϱ_{2Y} increases for $p = 20$ and decreases for $p = 6$ with an increase in γ.

Considerably more variation is obtained in the efficiencies of $\hat{Y}_1$, with all parameters β, γ and p having pronounced effects in certain circumstances. This predictor is the most efficient when $\gamma = 1$, which we knew from (23). The efficiency increases with γ, increases with p for large γ but decreases with p for small γ, and increases with β but β has less of an effect as p becomes larger and has no effect when $p = \infty$. The efficiency is appreciably greater than ϱ_{2Y} only for small β and no smaller than intermediate γ, and for large γ and no larger than intermediate β.

The set of coefficients for $\hat{Y}_3$ is in some respects intermediate to those for $\hat{Y}_2$ and $\hat{Y}_1$. For most situations in Tables 3 and 4, ϱ_{3Y} is intermediate to ϱ_{2Y} and ϱ_{1Y}. However, near the value of γ for which ϱ_{1Y} and ϱ_{2Y} switch relative magnitudes ϱ_{3Y} is less than both. This feature is demonstrated only for $p \geq 20$ because of rounding and of the large intervals used for γ.

It should be noted that there is a variety of combinations of γ and β for which all three efficiencies, ϱ_{1Y}, ϱ_{2Y} and ϱ_{3Y}, are fairly similar. These combinations include intermediate γ and intermediate to small β which may well prevail in many experiments. Also, although none of the three is the best, none of the three efficiencies is far below that of the best.

The efficiency, $\varrho_{\tilde{Y}Y}$, of selecting on the basis of double cross estimates, like ϱ_{2Y}, increases with β and p, and is affected little by variations in γ. Also, as we already knew, $\varrho_{\tilde{Y}Y}$ approaches one as β approaches one, since when $\sigma_\varepsilon^2/r = 0$ selection is among the true values of the double crosses. The surprising feature is that $\varrho_{\tilde{Y}Y}$ is never greater than ϱ_{2Y} and for many likely combinations of γ and β it is considerably less. Also, the other predictors, $\hat{Y}_1$ and $\hat{Y}_2$, are more efficient than $\tilde{Y}$ for most situations.

The standard deviation, σ_Y, among the true values behaves similar to ϱ_{2Y}. It is mainly affected by and increases with β. It increases with p and much more so for large γ than for small γ. It increases with γ for large p but the reverse happens for small p. This interaction of p and γ stems from a reduction in σ_Y with an increasing proportion of nonadditive variance, but which is offset for small p by the finite correcting coefficient of the additive variance (see 73

in the Appendix). When $p = \infty$ the variance is that among unrelated double crosses, *i. e.*, $\sigma^2_Y = \beta\,(1 + \gamma)/4$.

Discussion

A comparison of the fixed and random sample approach is pertinent to the remaining discussion. Let the model for the double crosses be

$$Y = \mu + G\,, \tag{45}$$

where G is the genetic effect. Let the model for two alternative single cross predictors be

$$\begin{aligned} X_1 &= \mu_1 + G + e_1\,, \\ X_2 &= \mu_1 + G + g + e_2\,. \end{aligned} \tag{46}$$

Provision is made for the experimental errors to be different. One of the predictors, X_1, is unbiased genetically and has a prediction error of

$$e_1 \tag{47}$$

with variance

$$E(e_1^2) = \sigma^2_{e_1}\,. \tag{48}$$

Any constants involving the means fall out in comparisons among the Y's and their corresponding X's, and they will be ignored. The other predictor, X_2, is biased genetically, g, and has a prediction error of

$$g + e_2 \tag{49}$$

with an average quadratic value of

$$E(g^2) + \sigma^2_{e_2}\,. \tag{50}$$

In both cases, random or fixed, the genotypes and environments are assumed uncorrelated.

From a fixed sample standpoint, it is a particular set of single and double crosses to which the procedure is to be applied. Being unable to evaluate each g in most cases, primary emphasis is placed on finding the least biased predictors with minimum experimental error variance. None of the single cross predictors is unbiased genetically if there is epistasis, as was shown by EBERHART (1964), and only $\hat{Y}_2$ is unbiased if there are only additive and dominance effects of genes.

Now suppose that the procedure is viewed as being used repeatedly for samples of single and double crosses. Then in the parametric population g will have a mean of 0 and variance of $E(g^2) = \sigma^2_g$ even though the parametric population may be a peculiar one consisting only of single and double crosses from highly selected inbred lines. Further, G will be distributed with a mean of 0 and variance, σ^2_G. For generality, let g and G be correlated with covariance, σ_{gG}. With this viewpoint, the prediction error variance is $\sigma^2_{e_1}$ for X_1 as before and $\sigma^2_g + \sigma^2_{e_2}$ for X_2. Now consider alternative predictors $c_1 X_1$ and $c_2 X_2$ where c_1 and c_2 are constants. The constants which minimize the prediction error variances,

$$\left.\begin{aligned} \sigma^2_{Y\cdot c_1 X_1} &= E\,[Y - \mu - c_1\,(X_1 - \mu_1)]^2 \\ &= (1 - c_1)^2\,\sigma^2_G + c_1^2\,\sigma^2_{e_1}\,, \\ \sigma^2_{Y\cdot c_2 X_2} &= E\,[Y - \mu - c_2\,(X_2 - \mu_1)]^2 \\ &= (1 - c_2)^2\,\sigma^2_G - 2\,c_2\,(1 - c_2)\,\sigma_{gG} \\ &\quad + c_2^2\,(\sigma^2_g + \sigma^2_e)\,, \end{aligned}\right\} \tag{51}$$

are the regression coefficients

$$\left.\begin{aligned} c_1^* &= B_{YX_1} = \frac{\sigma^2_G}{\sigma^2_G + \sigma^2_{e_1}} = \varrho_{G_1}\frac{\sigma_G}{\sigma_{X_1}}\,, \\ c_2^* &= B_{YX_2} = \frac{\sigma^2_G + \sigma_{gG}}{\sigma^2_G + 2\,\sigma_{gG} + \sigma^2_g + \sigma^2_{e_2}} = \varrho_{G_2}\frac{\sigma_G}{\sigma_{X_2}} \end{aligned}\right\} \tag{52}$$

and the prediction error variances are

$$\left.\begin{aligned} \sigma^2_{Y\cdot c_1^* X_1} &= \sigma^2_G\,(1 - \varrho^2_{G_1})\,, \\ \sigma^2_{Y\cdot c_2^* X_2} &= \sigma^2_G\,(1 - \varrho^2_{G_2})\,. \end{aligned}\right\} \tag{53}$$

The gains from selecting on the basis of the predictors, assuming linearity between the X' s and Y' s,

$$\left.\begin{aligned} \Delta_{X_1} &= k\,\frac{\sigma^2_G}{\sqrt{\sigma^2_G + \sigma^2_{e_1}}} = k\,c_1^*\,\sigma_{X_1} = \Delta_{c_1^* X_1} = k\,\varrho_{G_1}\sigma_G\,, \\ \Delta_{X_2} &= k\,\frac{\sigma^2_G + \sigma_{gG}}{\sqrt{\sigma^2_G + \sigma_{gG} + \sigma^2_g + \sigma^2_{e_2}}} = k\,c_2^*\,\sigma_{X_2} \\ &= \Delta_{c_2^* X_2} = k\,\varrho_{G_2}\sigma_G\,, \end{aligned}\right\} \tag{54}$$

are invariant of the constant multipliers. The prediction error variances when in their minimum form (53) rank exactly in reverse order to the gains (54). In other than minimum form, the prediction error variances may not rank the predictors the same as the gains, although they do for $c_1 = c_2 = 1$ when $\sigma_{gG} = 0$, which is a simpler situation.

The foregoing illustrates the connection between minimum prediction error variances and gains when applied to random samples. The extension to the simultaneous weighting of two or more kinds of information, for example $d_1 X_1 + d_2 X_2$, into a single predictor is straightforward. The constants, d_1^* and d_2^*, which maximize the gain also minimize the prediction error variance, although any other set of constants, $t\,d_1^*$ and $t\,d_2^*$, where t is a positive constant, will give the same gain. The maximization of gain is simply an adaptation of the selection index procedure given in considerable detail by HENDERSON (1963), who also pointed out many applications. The same principles are involved in weighting of information from individuals and families in making selections considered by LUSH (1947).

It is seen then that genetically biased predictors may have less prediction error variance than unbiased ones depending upon the bias variance, the covariance between the bias and the true values, and the experimental error variances. These factors are automatically accommodated in the relative weighting of the various single cross means. The number of parental lines, p, determining the number of single crosses in some of the means plays an important part in the relative weights and in the gains, a reflection of its joint effect on experimental variances and genetic covariances of hybrid means.

The choice among predictors and minimization of prediction error variance in random samples always depends upon a knowledge of the variance parameters, which are never known exactly. For single cross prediction these are σ_{ZZ} in (7) and $\underline{\sigma}_{ZY}$ in (8). If one is not willing to assume the hybrids to be members of a random mating population, then σ_{ZZ} and $\underline{\sigma}_{ZY}$ must be estimated directly from single crosses and double crosses. If the hybrids are assumed to be members of a random mating population, then estimates of σ_{ZZ} and $\underline{\sigma}_{ZY}$ can be constructed from estimates of genetic and environmental variances

from other types of experiments (COCKERHAM, 1963). In any case, estimates are involved. The problem of using estimated variances to construct the b's (WILLIAMS, 1962) is that they must be reliable, otherwise the predictor or index may be useless. However, the hybrid prediction framework is much more restricted than that for the general selection index considered by WILLIAMS. The weights, $\underline{b}^*$, cannot be negative and must conform to relative bounds (21). PATEL (1962), in a somewhat similar situation of selecting among single crosses, showed that estimated weights when restricted to fall within known bounds led to efficient selection.

The empirical evaluations were made only for additive and dominance variance and assuming the hybrids to be members of a noninbred population. To compare the predictors empirically with much generality in the epistatic model would be a large task. Epistasis, like dominance, reduces ϱ_{*Y}, and also $\varrho_{\tilde{Y}Y}$ and σ_Y except for small p, as compared to an additive genetic model. Consequently, the gains, $\Delta_* \propto \varrho_{*Y}\sigma_Y$, $\Delta_{\tilde{Y}} \propto \varrho_{\tilde{Y}Y}\sigma_Y$ and $\Delta_Y \propto \sigma_Y$, are all reduced by epistasis for a given proportion of the total variance that is environmental. The amount of reduction depends on the kinds and relative amounts of epistatic variance, and will increase with higher order epistasis and more with dominance types than with additive types.

One of the surprising outcomes is the efficiency of all the single cross predictors relative to double cross estimates in Tables 3 and 4. It was pointed out previously that this relative advantage decreases some with epistasis. Also, the most efficient procedure for selecting on the $\tilde{Y}$'s has not been used. For each double cross there are eight subdivisions of the double cross estimates (see the Appendix), each related differently to the double cross. The most efficient method is to weight the means of these subdivisions, or correspondingly appropriate linear functions of them, just as was done for the single crosses in predicting double crosses, and in the same way that PATEL (1962) did in selecting on single cross estimates for single crosses. The most efficient double cross predictor should always be superior to single cross prediction. In practice, however, one is concerned with the prediction of double crosses without having to obtain $p\,(p-1)\,(p-2)\,(p-3)/8$ estimates.

Whether to choose one of JENKINS' (1934) three predictors, $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_3$, or to estimate the best one, $\hat{Y}_*$, is probably unimportant in practice. The choice would be the closest of the three to the estimated one and most likely would give comparable results. This conclusion is based on the empirical evaluations in Tables 3 and 4, where the best of the three is only slightly less efficient than $\hat{Y}_*$ coupled with the fact that an estimated predictor is never as efficient as $\hat{Y}_*$.

JENKINS (1934) did not find a great deal of difference among the correlations of the three single cross predictors with double cross performance. He was estimating the following correlations,

$$\varrho_{1\tilde{Y}} = \varrho_{1Y}\varrho_{\tilde{Y}Y}, \quad \varrho_{2\tilde{Y}} = \varrho_{2Y}\varrho_{\tilde{Y}Y}, \quad \varrho_{3\tilde{Y}} = \varrho_{3Y}\varrho_{\tilde{Y}Y}. \tag{55}$$

One can see from the tabulations in Tables 3 and 4 that there are many situations for which the three are not very different, and not enough different to be distinguished from data. Some paradoxes are brought out in the Tables. The variance of an estimated $\varrho_{i\tilde{Y}}$ may be reduced by increasing p or increasing replications represented by increasing β in the Tables. However, there are genetic situations for which the three $\varrho_{i\tilde{Y}}$'s become more alike with an increase in p and/or β. Whether the increased similarity is more than offset by the decrease in variance of the estimates requires further exploration, but there is the suggestion of not being offset. Certainly, reasonably good data are required for distinguishing among the predictors. The double cross values should be well enough estimated that if another similar but environmentally independent set of estimates, $\tilde{Y}'$, were available, the correlation between the two,

$$\varrho_{\tilde{Y}\tilde{Y}'} = \varrho^2_{\tilde{Y}Y}, \tag{56}$$

would be reasonably high.

EBERHART (1964) and EBERHART et al. (1964) considered the use of three-way crosses in addition to single crosses in the prediction of double crosses. From a fixed sample viewpoint, they were principally concerned with minimum genetic biased predictors. The prediction model (4) can be augmented to include three-way cross information. Taking into account the different ways in which three-way crosses relate to double crosses will add at least six new variables, Z's, to the model (4). These additional variables can probably be formulated into an uncorrelated set as was done for the single crosses, although some will be correlated with the Z's pertaining to the single crosses. Minimization of the prediction error variance and maximization of the gain is accomplished in the same manner as was done for only single cross prediction. The inclusion of three-way cross information properly weighted will always improve the efficiency of prediction over the use of single crosses alone. The amount of improvement can be evaluated empirically for various parametric situations.

Only a single replicated experiment was considered for simplicity, but only slight modifications are needed to include environmental factors and genetic by environmental interactions. For an experiment at l environments for example, one replaces M_3/r and M_2/r in (6) with

$$\left.\begin{aligned} \frac{M_3}{r\,l} &= \frac{\sigma^2}{r\,l} + \frac{\sigma^2_{gE}}{l} + (p-2)\,\sigma^2_g, \\ \frac{M_2}{r\,l} &= \frac{\sigma^2}{r\,l} + \frac{\sigma^2_{sE}}{l} + \sigma^2_s, \end{aligned}\right\} \tag{57}$$

where σ^2_{gE} and σ^2_{sE} are components of variance for general and specific by environments. The gene effects and variances and covariances of relatives are always defined to be free of environments and of interactions of genes with environments, and any estimates of them should take this into account. A corresponding modification must be made in $\sigma^2_{\tilde{Y}}$ for the inclusion of different environments. An extension to multiple environmental classifications is straightforward, just more complicated.

Prediction and gain have been considered from an internal sample standpoint. By this is meant that only contrasts among the Y's, and as predictors only contrasts among the X's, have been considered. In such case the means, both of double and single crosses, cancel out whether from a random or fixed sample viewpoint. The question arises as to how to develop comparative predictors for separate samples of single crosses in different experiments. The means of the samples will differ both genetically and environmentally. In most cases the experiments are in different years and often in different places. Consequently, environmental differences are much more than those reflected by averages of internal environmental effects. One solution is to augment the prediction equation (4) with $b_4 Z_4$ where Z_4 is a deviation of an experimental mean of single crosses from the mean of all experiments. The relative weight, b_4, depends mainly on the proportions of genetic and environmental variances among experiments for which little information is generally available. If the environmental differences are relatively large then b_4 will be relatively small, and the comparisons are very nearly among internal predictions. Internal predictions are probably the best and safest solution in practice.

Having decided on a particular predictor,

$$\hat{Y} = b_1 (X_n - X_{\bar{p}}) + b_2 (X_{\bar{p}} - 2 X_l + X_0) + b_3 (X_{\bar{l}} - \bar{X}) , \quad (58)$$

the equation may be rearranged algebraically to simplify calculations. One alternative which is computationally simple is

$$\hat{Y} = b'_1 X_p + b'_2 X_{\bar{p}} + b'_3 X_{\bar{l}} + b'_4 \bar{X} , \quad (59)$$

where

$$\left.\begin{aligned} b'_1 &= \frac{-b_1}{2} , \quad b'_2 = \frac{b_1}{2} + b_2 \frac{(p-1)(p-2)}{(p-4)(p-5)} , \\ b'_3 &= b_3 - b_2 \frac{2(p-1)^2}{(p-4)(p-5)} , \\ b'_4 &= b_2 \frac{p(p-1)}{(p-4)(p-5)} - b_3 . \end{aligned}\right\} \quad (60)$$

The last term, $b'_4 \bar{X}$, is a constant and may be dropped in ranking the double crosses, and also one of the coefficients may be standardized to one. For example,

$$b''_1 X_p + X_{\bar{p}} + b''_3 X_{\bar{l}} , \quad (61)$$

where

$$b''_1 = \frac{b'_1}{b'_2} , \quad b''_3 = \frac{b'_3}{b'_2} , \quad (62)$$

will rank the double crosses the same as (58).

Appendix

Variances among Double Crosses

From (38) the experimental value for double cross $AB \cdot CD$, for example, is

$$\tilde{Y}_{AB \cdot CD} = Y_{AB \cdot CD} + \bar{\varepsilon}_{AB \cdot CD} , \quad (63)$$

and the experimental error is the mean of the plot errors for the double cross,

$$\bar{\varepsilon}_{AB \cdot CD} = \frac{\Sigma \varepsilon_{AB \cdot CD}}{r} \quad (64)$$

with variance

$$\sigma^2_{\bar{\varepsilon}} = \frac{\sigma^2_\varepsilon}{r} . \quad (65)$$

We shall use the following measure of variance,

$$\sigma^2_{\tilde{Y}} = E (\tilde{Y}_{AB \cdot CD} - \bar{\tilde{Y}})^2 , \quad (66)$$

where $\bar{\tilde{Y}}$ is the mean of all double crosses under consideration. This measure of variance brings into consideration the number of double crosses and the correlations among them in keeping with the treatment of the single cross predictors. The variance reduces to that among the true genotypic values of the double crosses by making $\sigma^2_\varepsilon = 0$.

Expanding (66),

$$E (\tilde{Y}_{AB \cdot CD} - \bar{\tilde{Y}})^2 = E (\tilde{Y}^2_{AB \cdot CD}) - 2 E (\tilde{Y}_{AB \cdot CD} \bar{\tilde{Y}}) + E(\bar{\tilde{Y}}^2) , \quad (67)$$

it can be shown that

$$E(\tilde{Y}_{AB \cdot CD} \bar{\tilde{Y}}) = E(\bar{\tilde{Y}}^2) , \quad (68)$$

so that

$$\sigma^2_{\tilde{Y}} = E(\tilde{Y}^2_{AB \cdot CD}) - E(\tilde{Y}_{AB \cdot CD} \bar{\tilde{Y}}) . \quad (69)$$

Evaluating the first expectation,

$$E(\tilde{Y}^2_{AB \cdot CD}) = \mu^2 + C_{AB \cdot CD, AB \cdot CD} + \sigma^2_\varepsilon , \quad (70)$$

where $C_{AB \cdot CD, AB \cdot CD}$ is the covariance of the genotypic values of the double crosses with themselves, which is the variance among the true values of unrelated double crosses, but will be left in covariance terminology in accord with the remaining expectations. Turning now to the second expectation in (69), taking into account all possible types of related hybrids in the mean of the $p(p-1)(p-2) \times (p-3)/8$ double crosses, and assuming of course that all experimental errors are uncorrelated in any way, the following result is obtained

$$\begin{aligned} E(\tilde{Y}_{AB \cdot CD} \bar{\tilde{Y}}) = \mu^2 + \Big[& \sigma^2_\varepsilon + 1 && C_{AB \cdot CD, AB \cdot CD} \\ & + 2 && C_{AB \cdot CD, AC \cdot BD} \\ & + 4(p-4) && C_{AB \cdot C-, AB \cdot C-} \\ & + 8(p-4) && C_{AB \cdot C-, AC \cdot B-} \\ & + 4(p-4)(p-5) && C_{A-\cdot B-, A-\cdot B-} \\ & + 4(p-4)(p-5) && C_{A-\cdot B-, AB \cdot --} \\ & + (p-4)(p-5) && C_{AB \cdot --, AB \cdot --} \\ & + 2(p-4)(p-5)(p-6) && C_{A-\cdot --, A-\cdot --} \\ & + \frac{(p-4)(p-5)(p-6)(p-7)}{8} && C_{--\cdot --, --\cdot --} \Big] / \\ & \frac{p(p-1)(p-2)(p-3)}{8} . \end{aligned} \quad (71)$$

The designations of the covariances show the number of parental lines common and the manner in which they are common. The coefficients of the covariances add to the total number of double crosses, the last covariance with no lines common being included for completeness and is always zero. More details are given by COCKERHAM (1961) and RAWLINGS and COCKERHAM (1962). With the assumptions generally required in expressing the covariances of relatives

in terms of genetic variances (16) the additive and dominance coefficients are as follows:

$$\begin{array}{lcc} & \underline{\alpha} & \underline{\delta} \\ C_{AB\cdot CD,\,AB\cdot CD} & \frac{1}{2} & \frac{1}{4} \\ C_{AB\cdot CD,\,AC\cdot BD} & \frac{1}{2} & \frac{1}{8} \\ C_{AB\cdot C-,\,AB\cdot C-} & \frac{3}{8} & \frac{1}{8} \\ C_{AB\cdot C-,\,AC\cdot B-} & \frac{3}{8} & \frac{1}{16} \\ C_{A-\cdot B-,\,A-\cdot B-} & \frac{1}{4} & \frac{1}{16} \\ C_{A-\cdot B-,\,AB\cdot --} & \frac{1}{4} & 0 \\ C_{AB\cdot --,\,AB\cdot --} & \frac{1}{4} & 0 \\ C_{A-\cdot --,\,A-\cdot --} & \frac{1}{8} & 0 \\ C_{---\cdot --,\,---\cdot --} & 0 & 0 \end{array} \qquad (72)$$

Putting (70) and (71) together as in (69) and substituting the coefficients (72) into the covariances as in (16) leads to the appropriate expressions for $\sigma^2_{\bar{Y}}$ and σ^2_Y,

$$\left.\begin{aligned} \sigma^2_{\bar{Y}} &= \left(1 - \frac{8}{p\,(p-1)\,(p-2)\,(p-3)}\right)\frac{\sigma^2_\varepsilon}{r} + \sigma^2_Y\,, \\ \sigma^2_Y &= \left(1 - \frac{4}{p}\right)\frac{\sigma^2_A}{2} + \left(1 - \frac{8}{p\,(p-1)}\right)\frac{\sigma^2_D}{4} \\ &\quad + \left(1 - \frac{p+8}{p\,(p-1)}\right)\frac{\sigma^2_{AA}}{4} + \cdots . \end{aligned}\right\} \qquad (73)$$

The error variance, σ^2_ε, is not exactly comparable to that, σ^2, for the single crosses. The error variance for single crosses, in the absence of competitive effects, can be decomposed into

$$\sigma^2 = \frac{\sigma^2_1}{n} + \sigma^2_2\,, \qquad (74)$$

where σ^2_1 is the environmental variance among n plants in a plot and σ^2_2 is the plot component of environmental variance. In contrast,

$$\sigma^2_\varepsilon = \frac{\sigma^2_1 + C_{S2} - C_{AB\cdot CD,\,ABCD}}{n} + \sigma^2_2\,, \qquad (75)$$

where $C_{S2} - C_{AB\cdot CD,\,ABCD}$ represents the genotypic variance among individuals of the same double cross in a plot, and C_{S2}, given in (17), represents the total genotypic variance.

Zusammenfassung

Für die Vorausschätzung von Doppelkreuzungsbastarden aus Einzelkreuzungshybriden wird eine einheitliche Theorie entwickelt, die sowohl genetische als auch experimentelle Bedingungen berücksichtigt. Die Methode ist der für die Berechnung von Selektionsindizes analog. Es wird die Beziehung des Vorhersagemodells zum genetischen Modell erläutert. JENKINS' (1934) drei Einzelkreuzungs-Schätzwerte, der beste Einzelkreuzungs-Schätzwert und die Selektion auf der Grundlage der Doppelkreuzungs-Schätzungen werden für ein additives und Dominanz-Modell mit variierenden Verhältnissen der experimentellen Fehlervarianz und unterschiedlicher Anzahl von Bastarden empirisch miteinander verglichen. Der Unterschied zwischen fixierter und zufälliger genetischer Stichprobenmethode wird hinsichtlich der Vorhersage besprochen.

References

1. COCKERHAM, C. CLARK: Implications of genetic variances in a hybrid breeding program. Crop Science **1**, 47–52 (1961). — 2. COCKERHAM, C. CLARK: Estimation of genetic variances. Statistical Genetics and Plant Breeding. National Academy of Sciences-National Research Council Publ. 982, 53–94 (1963). — 3. COCKERHAM, C. CLARK: Group inbreeding and coancestry. (Submitted to Genetics) (1967). — 4. EBERHART, S. A.: Theoretical relations among single, three-way, and double cross hybrids. Biometrics **20**, 522–539 (1964). — 5. EBERHART, S. A., W. A. RUSSELL, and L. H. PENNY: Double cross hybrid prediction in maize when epistasis is present. Crop Science **4**, 363–366 (1964). — 6. HENDERSON, C. R.: Selection index and expected genetic advance. Statistical Genetics and Plant Breeding. National Academy of Sciences-National Research Council Publ. 982, 141–163 (1963). — 7. JENKINS, MERLE T.: Methods of estimating the performance of double crosses in corn. J. Amer. Soc. Agronomy **26**, 199–204 (1934). — 8. LUSH, JAY L.: Family merit and individual merit as bases for selection. Amer. Naturalist **81**, 241–261; 362–379 (1947). — 9. PATEL, R. M.: Selection among factorially classified variables. Ph. D. Thesis, North Carolina State University at Raleigh (1962). — 10. RAWLINGS, J. O., and C. CLARK COCKERHAM: Analysis of double cross hybrid populations. Biometrics **18**, 229–244 (1962). — 11. SPRAGUE, G. F.: Chapter V: Corn breeding. In: Corn and Corn Improvement. New York, N.Y.: Academic Press Inc. 1955. — 12. WILLIAMS, J. S.: Some statistical properties of a genetic selection index. Biometrika **49**, 325–337 (1962).

A Comprehensive Breeding System*

S. A. EBERHART, M. N. HARRISON and **F. OGADA** [1]

Maize Research Section, Kitale, Kenya

Summary. An outline of a comprehensive breeding system developed and used by the Kenya Maize Research Section is presented. This system has four main phases:

1. Evaluation of local and exotic varieties so that the breeding material is the best available.

2. Compositing the selected breeding material into two or more populations or varieties in such a manner that each population has considerable genetic variation for the traits requiring improvement and that the crosses of these populations will show heterosis.

3. Recurrent selection in each population to increase the frequency of favorable genes so that the populations and population crosses are improved with each cycle of selection.

4. Release of a commercial variety of one of the following forms: (a) the cross of two populations as a variety cross hybrid; (b) single, three-way or double cross hybrids from inbred lines developed from the elite material after each cycle of selection; or (c) a synthetic variety derived from the advanced generation of the population cross in areas where hybrid production is not yet feasible.

Preliminary results are presented to indicate the improvement possible in maize by use of this system. Its possible extension to other crops is also briefly discussed.

* Dedicated to Dr. GEORGE F. SPRAGUE on the occasion of his 65th birthday.

[1] Research Geneticist, Agricultural Research Service, U.S. Department of Agriculture/U.S. Agency for International Development/East African Agriculture & Forestry Research Organisation; Senior Maize Research Officer and Maize Breeder, Maize Research Section, Kenya Ministry of Agriculture.