

24

Applications of Index Selection

Draft Version 4 August 1997, ©Dec. 2000, B. Walsh and M. Lynch

Please email any comments/corrections to: jbwalsh@u.arizona.edu

Improving the Response of a Single Character Using a Selection Index

With univariate selection, the only information used to predict response is the single-character phenotypic value of an individual. Often considerably more information relevant to that character is available, such as repeated measures of that character over time, correlated characters in the same individual, or values from relatives. Incorporating this information into a Smith-Hazel index improves the response over that of simple univariate selection on the character, as suggested by Hazel (1943) and further developed by numerous authors such as Lush (1944, 1947), Rendel (1954), Osborne (1957a,b,c), Purser (1960), Searle (1965) and Gjedrem (1967a,b).

Our discussion of using selection indices to improve response in a single character covers the last three sections of this chapter. This section develops the general theory (which follows as a simplification of the Smith-Hazel index) and then applies these results to three important cases — a general analysis when either phenotypic or genotypic correlations are zero, improving response using repeated measurements of a characters over time, and selection on a ratio. The next section examines the use of information from relatives to improve response with a special focus on combined selection (the optimal weighting of individual and family information), while the final section considers marker-assisted selection.

General theory. All the results of the Smith-Hazel index apply, but when our interest is the response of only a single character considerable simplification occurs in many of the results. Let z_1 be the character of interest (the *primary character*) and z_2, \dots, z_n be $n - 1$ other *secondary characters* that potentially provide information on the primary character. Since the only response of the primary character is of interest, the vector of economic weights a has $a_1 = 1$ and all other elements zero.

Writing the additive-genetic variance-covariance matrix as $\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n)$ where $\mathbf{g}_i^T = (g_{1i}, g_{2i}, \dots, g_{ni})$ is the vector of additive genetic covariances between character i and all other characters, we have

$$\mathbf{G}\mathbf{a} = (\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_n) \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathbf{g}_1$$

where \mathbf{g}_1 is the vector of additive genetic covariances of the primary character with all other characters being considered. For notational ease, we drop the subscript and simply use \mathbf{g} and likewise use g for the additive genetic value of character 1. Applying Equation 15.11, the vector of weights for the Smith-Hazel index simplifies to

$$\mathbf{b}_s = \mathbf{P}^{-1}\mathbf{G}\mathbf{a} = \mathbf{P}^{-1}\mathbf{g} \quad (24.29a)$$

giving the index as

$$I_s = \mathbf{b}_s^T \mathbf{z} = \mathbf{g}^T \mathbf{P}^{-1} \mathbf{z} \quad (24.29b)$$

Substituting $\mathbf{G}\mathbf{a} = \mathbf{g}$ into Equation 15.12 gives the response as

$$R = \bar{i} \cdot \sqrt{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}} \quad (24.29c)$$

Under univariate selection, $R = \bar{i} \cdot h_1^2 \sigma_{z_1} = \bar{i} \cdot h_1 \sigma_g$, giving the increase in response using index selection as

$$\frac{\sqrt{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}}}{h_1 \sigma_g} \quad (24.29d)$$

An alternative way to quantify the advantages of an index over univariate selection is to consider how much variation in g is accounted for by the index. Since the correlation between an individual's phenotypic and additive-genetic values is $\rho_{g, z_1} = \sigma_{g, z_1} / \sigma_g \sigma_{z_1} = \sigma_g^2 / \sigma_{z_1}^2 = h_1$, the accuracy of using only z_1 to predict g is $\rho_{g, z_1}^2 = h_1^2$. From Equation 15.14b, the best linear predictor of the breeding value of the primary character ($g - \mu$) is

$$\mathbf{b}_s^T (\mathbf{z} - \boldsymbol{\mu}) = \mathbf{g}^T \mathbf{P}^{-1} (\mathbf{z} - \boldsymbol{\mu}) = I_s - \mathbf{g}^T \mathbf{P}^{-1} \boldsymbol{\mu} \quad (24.30a)$$

which is I_s when the \mathbf{z} rescaled to have zero mean. The accuracy of this index in predicting g is

$$\rho_{H, I_s}^2 = \frac{\sigma_{H, I_s}^2}{\sigma_H^2 \cdot \sigma_{I_s}^2} = \frac{(\mathbf{b}_s^T \mathbf{g})^2}{\sigma_g^2 \cdot \mathbf{b}_s^T \mathbf{P} \mathbf{b}_s} = \frac{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}}{\sigma_g^2} \quad (24.30b)$$

so that the improvement in accuracy by using an index is

$$\frac{\rho_{g, I_s}^2}{\rho_{g, z_1}^2} = \frac{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}}{h_1^2 \cdot \sigma_g^2}$$

Equations 24.29 and 30 give the general expressions for improving response in a single character using selection indices and can be applied to a very wide variety of situations.

Example 4. Robinson et al. (1951) estimated the genotypic and phenotypic covariances between yield and several other characters in maize. Using their estimates, construct the optimal index to improve yield (z_1) using plant height (z_2) and ears per plant (z_3) as secondary characters. The estimated phenotypic covariance matrix for these characters is

$$\hat{\mathbf{P}} = \begin{pmatrix} 0.0069 & 0.0968 & 0.0132 \\ 0.0968 & 28.8796 & 0.2313 \\ 0.0132 & 0.2313 & 0.0526 \end{pmatrix}$$

while the vector of estimated additive genotypic covariances between yield and other characters is

$$\hat{\mathbf{g}} = \begin{pmatrix} 0.0028 \\ 0.0964 \\ 0.0075 \end{pmatrix}$$

For yield, $\sigma_g^2 = 0.0028$ and $h_1^2 = 0.0028/0.0069 \simeq 0.41$, giving $h_1\sigma_g = \sqrt{0.0028 \cdot 0.41} \simeq 0.0339$, and an expected response to selection solely on yield as $R = 0.0339 \cdot \bar{z}$. The optimal index incorporating both yield and the two secondary characters is

$$I_s = \hat{\mathbf{g}}^T \hat{\mathbf{P}}^{-1} \mathbf{z} = 0.23 \cdot z_1 + 0.002 \cdot z_2 + 0.075 \cdot z_3$$

which has expected response

$$\bar{z} \sqrt{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}} = \bar{z} \sqrt{0.00141} = \bar{z} \cdot 0.0375$$

an 11 percent increase relative to selection on yield only. The accuracy of this index is $\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g} / \sigma_g^2 = 0.00141 / 0.0028 \simeq 0.504$, so that I_s accounts for 50.4 percent of the additive genetic variance in yield, while the phenotype of yield alone accounts for only $h^2 = 0.41$, or 41 percent. Increasing yield is a common use of an indirect index. However, experiments reviewed by Pritchard et al. (1973) shows that usually the index is only slightly better than direct selection and often can be worse (likely due to sampling errors giving the estimated index incorrect weights). Index selection is most superior when environmental effects overwhelm genetic differences.

A key concern in constructing an index is which secondary characters to include. If \mathbf{g} and \mathbf{P} are estimated without error, addition of any correlated (genetic

or phenotypic) character always increases the accuracy of the index. However, genetic parameters are estimated with error and the inclusion of characters that are uncorrelated, but show a estimated correlation due to sampling effects, reduces the efficiency of the index. Sales and Hill (1976a) find that the greatest errors occur when the primary character has low heritability, but this is exactly the case where a selection index is potentially the most useful (Gjedrem 1967a). Bouchez and Goffinet (1990) suggest a robust procedure for evaluating which secondary characters to exclude.

***More detailed analysis of two special cases.** First suppose there are no phenotypic correlations between the characters so that \mathbf{P} (and hence \mathbf{P}^{-1}) is diagonal. In this case, the i th diagonal element of \mathbf{P}^{-1} is $1/P_{ii} = 1/\sigma_{z_i}^2$, giving

$$\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g} = \sum_{j=1}^n \frac{[\sigma(g, g_j)]^2}{\sigma_{z_j}^2} = \left(\frac{\sigma_g^4}{\sigma_{z_1}^2} \right) \left(1 + \sum_{j=2}^n \frac{[\sigma(g, g_j)]^2 \sigma_{z_1}^2}{\sigma_g^4 \sigma_{z_j}^2} \right)$$

Using $\sigma(g, g_j) = \rho_j \sigma_g \sigma_{g_j}$ where ρ_j is the correlation between additive genetic values of character j and the primary character, the response can be expressed as

$$R = \bar{v} h_1 \sigma_g \sqrt{1 + \frac{1}{h_1^2} \sum_{j=2}^n \rho_j^2 h_j^2} \quad (24.31a)$$

Hence the increase in response in z_1 using an index is

$$\sqrt{1 + \frac{1}{h_1^2} \sum_{j=2}^n \rho_j^2 h_j^2} \quad (24.31b)$$

This is strictly greater than one unless z_1 is genetically uncorrelated with all the other considered characters in which case it equals one. The advantage of index selection increases as either the heritabilities of correlated characters increase or as the heritability of z_1 decreases. Thus, when the heritability of z_1 is low using an index can result in a significantly increased response.

The second case is when none of the secondary characters are genetically correlated with the primary character. Rendel (1954) considered this as a means of using a second character to increase the heritability of the first. Rendel's idea is that a second phenotypically correlated character potentially provides information on the environmental value of the primary character, reducing uncertainty as to its genotypic value and as a consequence increasing heritability. Here $\mathbf{g} = \sigma_g^2(1, 0, \dots, 0)^T$ implying

$$\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g} = \sigma_g^4 (1 \quad 0 \quad \dots \quad 0) \mathbf{P}^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \sigma_g^4 P_{11}^{-1}$$

where P_{11}^{-1} denotes the 1, 1 element of \mathbf{P}^{-1} . Substituting into Equation 24.29c gives the response as

$$R = \bar{i} \sigma_g^2 \sqrt{P_{11}^{-1}} = \bar{i} h_1 \sigma_g \sqrt{\sigma_{z_1}^2 P_{11}^{-1}} \tag{24.32}$$

and hence the increase in response using an index is $\sqrt{\sigma_{z_1}^2 P_{11}^{-1}}$. To see that this expression is greater than or equal to one, we first digress on a useful identity from matrix algebra. Partitioning the phenotypic variance-covariance matrix as

$$\mathbf{P} = \begin{pmatrix} P_{11} & \mathbf{p}^T \\ \mathbf{p} & \mathbf{S} \end{pmatrix} \quad \text{where} \quad \mathbf{p} = \begin{pmatrix} P_{12} \\ \vdots \\ P_{1n} \end{pmatrix} \quad \text{and} \quad \mathbf{S} = \begin{pmatrix} P_{22} & \cdots & P_{2n} \\ \vdots & \ddots & \vdots \\ P_{n2} & \cdots & P_{nn} \end{pmatrix}$$

following Cunningham (1969), it can be shown that

$$P_{11}^{-1} = (P_{11} - \mathbf{p}^T \mathbf{S}^{-1} \mathbf{p})^{-1} = \sigma_{z_1}^{-2} \left(1 - \frac{\mathbf{p}^T \mathbf{S}^{-1} \mathbf{p}}{\sigma_{z_1}^2} \right)^{-1} \tag{24.33}$$

giving the response as

$$R = \bar{i} h_1 \sigma_g \left(1 - \frac{\mathbf{p}^T \mathbf{S}^{-1} \mathbf{p}}{\sigma_{z_1}^2} \right)^{-1/2} \tag{24.34a}$$

showing that increase in response using an index is

$$\left(1 - \frac{\mathbf{p}^T \mathbf{S}^{-1} \mathbf{p}}{\sigma_{z_1}^2} \right)^{-1/2} \tag{24.34b}$$

which is greater than one if the quadratic product term is positive. Since \mathbf{S} is itself a covariance matrix, it is positive-definite (unless one of the secondary characters can be expressed as a linear combination of the others in which case it is non-negative definite). Recall from Chapter 31 that if \mathbf{S} is positive definite, so is \mathbf{S}^{-1} and hence the quadratic product $\mathbf{p}^T \mathbf{S}^{-1} \mathbf{p} > 0$ unless $\mathbf{p} = \mathbf{0}$. This later case occurs when z_1 is phenotypically uncorrelated with all secondary characters being considered, in which case index selection gives the same response as univariate selection.

Repeated measures of a character. Suppose a single character is measured at n different times to give a vector of observations (z_1, \dots, z_n) for each individual. Under what conditions does the use of such *repeated measures* improve response? The idea is that if some of the environmental effects change from one measurement

to the next, multiple measurements average out these effects. The simplest model of environmental effects is that the j th measurement can be decomposed as $z_j = g + e_p + e_j^*$ where e_p the permanent environmental effects (which also includes non-additive genetic terms if they are present) and e_j^* is the transient part of the environment which is assumed to be uncorrelated from one measurement to the next. Thus

$$\sigma(z_k, z_j) = \sigma(g + e_p + e_k^*, g + e_p + e_j^*) = \begin{cases} \sigma_z^2 & \text{for } k = j \\ \sigma_g^2 + \sigma_e^2 \equiv t \cdot \sigma_z^2 & \text{for } k \neq j \end{cases}$$

where $t = (\sigma_g^2 + \sigma_{e_p}^2) / \sigma_z^2$ is correlation between measurements, previously defined as the *repeatability* of the character (Chapter 5). The covariance in additive genetic values between measurements is $\sigma(g_i, g_j) = \sigma(g, g) = \sigma_g^2$. Hence,

$$\mathbf{g} = \sigma_g^2 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{P} = \sigma_z^2 \begin{pmatrix} 1 & t & \cdots & t \\ t & 1 & \cdots & t \\ \vdots & \vdots & \ddots & \vdots \\ t & t & \cdots & 1 \end{pmatrix}$$

To compute the vector of weights \mathbf{b}_s , first note the following identity: for the $m \times m$ matrix

$$\mathbf{A} = \begin{pmatrix} 1 & a & \cdots & a \\ a & 1 & \cdots & a \\ \vdots & \vdots & \ddots & \vdots \\ a & a & \cdots & 1 \end{pmatrix}, \quad \text{then} \quad \mathbf{A}_{ij}^{-1} = \begin{cases} \frac{1 + (m-2)a}{1 + (m-2)a - (m-1)a^2} & \text{for } i = j \\ \frac{-a}{1 + (m-2)a - (m-1)a^2} & \text{for } i \neq j \end{cases} \quad (24.35)$$

Using this identity, a little algebra gives

$$\mathbf{b}_s = \mathbf{P}^{-1} \mathbf{g} = \frac{h^2}{1 + t(n-1)} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad (24.36)$$

Noting that $I_s = \mathbf{b}_s^T \mathbf{z} = c \cdot \bar{z}$, the index can be rescaled to simply \bar{z}_i (the average of all measurements for an individual). Since

$$\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g} = \left(\frac{h^2 \sigma_g^2}{1 + t(n-1)} \right) (1 \quad \cdots \quad 1) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = h^2 \sigma_g^2 \left(\frac{n}{1 + t(n-1)} \right)$$

the accuracy of this index in predicting additive-genetic values is

$$\rho_{I,g}^2 = \frac{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}}{\sigma_g^2} = h^2 \cdot \left(\frac{n}{1 + t(n-1)} \right) \quad (24.37a)$$

with resulting response to selection

$$R = \bar{i} \cdot h_1 \sigma_g \sqrt{\frac{n}{1 + t(n - 1)}} \tag{24.37b}$$

as obtained by Berge (1934). The ratio of response under the index to response using a single measurement approaches $t^{-1/2}$ for large n , so that for repeatabilities of $t = 0.1, 0.25, 0.5,$ and $0.75,$ it approaches 3.2, 2, 1.4, and 1.2. Significant gain in response can occur if repeatability is low, while there is little advantage when repeatability is high. Balancing any potential gain in response is an increase in cost and potentially longer breeding time (Turner and Young 1969). More generally, repeated measures can be modified to allow for correlations between transient environment effects by suitably modifying \mathbf{P} . In such cases, the index may weight separate measurements differentially so that the index can be significantly different from the simple average value of repeated measures.

Selection on a ratio. Occasionally, it is desirable to select on a ratio of two measured characters. Feed efficiency, defined as the ratio of feed intake to growth rate, is a classic example from animal breeding. Let $r = z_1/z_2$ be the desired ratio based on characters z_1 and z_2 . While approximations for the response to selection on r have been developed (Turner 1959, Hühn 1992), direct selection on r is less efficient than selection on an index based on z_1 and z_2 (Gunsett 1984, 1986, 1987, Mather et al. 1988, Campo and Rodriguez 1990). Since the merit function $H = z_1/z_2$ is non-linear, different approaches for constructing the optimal linear index have been proposed. Turner (1959) suggested taking logs to give the linear index $I = y_1 - y_2$ using the new characters y_1 and y_2 where $y_i = \ln(z_i)$ and gives expressions for the heritability of a ratio in this case. The downside to this approach is that it requires obtaining estimates of the phenotypic and additive-genetic covariance matrices for the transformed vector \mathbf{y} . Alternatively, Lin (1980) and Gunsett (1984) suggest using the vector of untransformed variables \mathbf{z} with economic weights given by approximating H by a first-order Taylor series (Equation 15.27). Since

$$\left. \frac{\partial z_1/z_2}{\partial z_1} \right|_{z_1=\mu_1} = \frac{1}{\mu_2}, \quad \text{and} \quad \left. \frac{\partial z_1/z_2}{\partial z_2} \right|_{z_2=\mu_2} = \frac{-\mu_1}{\mu_2^2}$$

the vector of economic weights becomes

$$\mu_2 \cdot \mathbf{a} = \begin{pmatrix} 1 \\ -\mu_1/\mu_2 \end{pmatrix} \tag{24.38}$$

where μ_i is the current mean of character i , so that the economic weights change each generation to reflect changes in character means. The Smith-Hazel index each generation is constructed by using either $\mathbf{a}^T = (1, -\mu_1/\mu_2)$ or $\mathbf{a}^T = (\mu_2, -\mu_1)$.

An alternative approach considered by Famula (1990) and Campo and Rodriguez (1990) is to construct a nonlinear index consisting of the ratio of two linear indices,

$$\frac{I_1}{I_2} = \frac{\mathbf{b}_1^T(\mathbf{z} - \boldsymbol{\mu})}{\mathbf{b}_2^T(\mathbf{z} - \boldsymbol{\mu})} \quad (24.39a)$$

where I_i is the index that gives the best linear predictor of the breeding value of character i using information from both z_1 and z_2 . This should be an improvement in predicting the value of g_i over that predicted just using the value of z_i alone unless both characters are phenotypically and genetically uncorrelated. Applying Equation 24.30a, the weights on these indices are given by

$$\mathbf{b}_1 = \begin{pmatrix} \sigma_{z_1}^2 & \sigma_{z_1, z_2} \\ \sigma_{z_1, z_2} & \sigma_{z_2}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{g_1}^2 \\ \sigma_{g_1, g_2} \end{pmatrix}, \quad \text{and} \quad \mathbf{b}_2 = \begin{pmatrix} \sigma_{z_1}^2 & \sigma_{z_1, z_2} \\ \sigma_{z_1, z_2} & \sigma_{z_2}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{g_1, g_2} \\ \sigma_{g_2}^2 \end{pmatrix} \quad (24.39b)$$

The difference between this non-linear index and the Smith-Hazel index constructed using Equation 24.38 is that the latter attempts to predict the ratio directly, while the nonlinear index attempts to predict the denominator and numerator separately.

There is uncertainty as to which method is most efficient. While Famula (1990) showed theoretically that there should be very little difference between the two, this was not observed by Campo and Rodriguez (1990), who selected for increased values of egg mass/ adult weight ratios in *Tribolium castaneum*. They found that this ratio did not respond to direct selection (response after three generations was $R = 0.82 \pm 1.56$). Selection on a linear index with economic weights given by Equation 24.38 was effective ($R = 1.92 \pm 0.44$), while the greatest response was observed by selection on the nonlinear index given by Equation 24.39a ($R = 5.94 \pm 1.52$). Hence while selection using either index was more efficient than selection directly on the character, the nonlinear index produced the largest response.

Using Information From Relatives

Often measurements of the character of interest exist for relatives and this information can easily be incorporated into a selection index to both improve response and increase the accuracy of predicted breeding values. We mention in passing here that although our discussion is restricted to the case where the character of interest is the one measured in relatives, other measured characters in relatives could also be incorporated using standard index theory. The basic theory presented here very generally extends to arbitrary sets of relatives, although more powerful methods for estimating breeding values (BLUP and REML, reviewed in

Chapter 20) exist. We start by reviewing the general theory and then examining family selection in detail.

General Theory. Since our interest is response in a single character, we build upon the simplifications of the Smith-Hazel index developed in the previous section. One significant difference with using relatives to construct an index is that far fewer parameters have to be estimated. A general index with n secondary characters has $(n + 1)(n + 4)/2$ parameters to estimate ($(n + 1)(n + 2)/2$ phenotypic covariances and $n + 1$ additive-genetic covariances). However, if the index uses measures from known relatives then only the significant variance components for the character need be estimated, as the elements of \mathbf{G} and \mathbf{P} can then be constructed from the theory of correlation between relatives (Chapter 6). For example, if non-additive genetic variance is not significant and genotype-environment interactions and maternal effects can be ignored, only σ_z^2 and h^2 need be estimated regardless of how many relatives are measured.

Let z_1 denote the character value measured in the individual of interest and z_2, \dots, z_{n+1} be measurements of this character in n of its relatives. Since we are only interested in the response in z_1 , then (similar to the previous section) only the vector of additive genetic covariances \mathbf{g} between the individual of interest and each of its relatives is required. The selection index ignores additional information from the full additive-genetic covariance matrix \mathbf{G} , such as the genetic covariances between measured relatives. The method of BLUP incorporates this additional information (the covariances between *all* sets of relatives) and not surprisingly has a greater accuracy in predicting breeding value than a selection index. Under the assumption that the character has the same phenotypic and additive genetic variance in all relatives, it is useful to work with correlations, rather than covariances. Denote by \mathbf{P}_ρ the matrix of phenotypic correlations between characters, hence

$$\mathbf{P} = \begin{pmatrix} \sigma(z_1, z_1) & \cdots & \sigma(z_1, z_{n+1}) \\ \vdots & \ddots & \vdots \\ \sigma(z_{n+1}, z_1) & \cdots & \sigma(z_{n+1}, z_{n+1}) \end{pmatrix} = \sigma_z^2 \cdot \begin{pmatrix} 1 & \cdots & \rho_{z_1, z_{n+1}} \\ \vdots & \ddots & \vdots \\ \rho_{z_{n+1}, z_1} & \cdots & 1 \end{pmatrix} = \sigma_z^2 \cdot \mathbf{P}_\rho \quad (24.40a)$$

Likewise, let \mathbf{g}_ρ denote the vector of additive-genetic correlations between the individual of interest and its relatives, with

$$\mathbf{g} = \begin{pmatrix} \sigma(g, g) \\ \sigma(g, g_2) \\ \vdots \\ \sigma(g, g_{n+1}) \end{pmatrix} = h^2 \sigma_z^2 \cdot \begin{pmatrix} 1 \\ \rho_{g, g_2} \\ \vdots \\ \rho_{g, g_{n+1}} \end{pmatrix} = h^2 \sigma_z^2 \cdot \mathbf{g}_\rho \quad (24.40b)$$

From Equation 24.29a the resulting Smith-Hazel index weights are

$$\mathbf{b}_s = \mathbf{P}^{-1} \mathbf{g} = h^2 \cdot \mathbf{P}_\rho^{-1} \mathbf{g}_\rho \quad (24.41a)$$

giving (from Equation 24.30a) the best linear predictor of the breeding value for the individual of interest as

$$\mathbf{b}_s^T (\mathbf{z} - \boldsymbol{\mu}) = h^2 \cdot \mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} (\mathbf{z} - \boldsymbol{\mu}) \quad (24.41b)$$

From Equation 24.30b the accuracy of this index in predicting breeding value is

$$\rho_{g,I}^2 = \frac{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}}{\sigma_g^2} = h^2 \cdot \mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho \quad (24.41c)$$

Since the accuracy in predicting breeding value from a single measure of an individual's phenotype is h^2 , the increase in accuracy using information from relatives is given by the quadratic product $\mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho$. From Equation 24.29c the expected response to selection on this index is

$$\frac{R}{\bar{i}} = \sqrt{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}} = h \sigma_g \cdot \sqrt{\mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho} \quad (24.41d)$$

Information from a single relative. As our first application, consider the simplest case of a single measurement from an individual z_1 and a single relative z_2 . Letting ρ_p and ρ_g be the phenotypic and additive-genetic correlations between the individual and this relative, we have

$$\mathbf{g}_\rho = \begin{pmatrix} 1 \\ \rho_g \end{pmatrix}, \quad \mathbf{P}_\rho = \begin{pmatrix} 1 & \rho_z \\ \rho_z & 1 \end{pmatrix} \quad \text{hence} \quad \mathbf{P}_\rho^{-1} = (1 - \rho_z^2)^{-1} \begin{pmatrix} 1 & -\rho_z \\ -\rho_z & 1 \end{pmatrix} \quad \blacksquare$$

Applying 24.41a, and rescaling the index so that the weight on z_1 is one gives the Smith-Hazel index for this situation as

$$I_s = z_1 + \left(\frac{\rho_g - \rho_z}{1 - \rho_z \rho_g} \right) z_2 \quad (24.42a)$$

Similarly,

$$\begin{aligned} \mathbf{g}^T \mathbf{P}^{-1} \mathbf{g} &= h^2 \sigma_g^2 \mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho \\ &= h^2 \sigma_g^2 \left(1 + \frac{(\rho_g - \rho_z)^2}{1 - \rho_z^2} \right) \end{aligned} \quad (24.42b)$$

so that the increase in response over univariate selection is

$$\frac{\sqrt{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}}}{h \cdot \sigma_g} = \sqrt{1 + \frac{(\rho_g - \rho_z)^2}{1 - \rho_z^2}} \quad (24.42c)$$

Constructing selection indices when the individual itself is not measured. An important class of applications is the construction of indices to predict breeding value when the individual is not (or cannot be) measured. For example, consider a female-limited character. Selection on males would increase response but the character cannot be scored. Information from females relatives can be used to construct an index to predict breeding value in males, and hence allow for selection in males. Another example is when an individual must be sacrificed to measure the character, such as selection on internal organs. In such cases information from scored sibs can be used to predict breeding value.

Indirect indices wherein the primary character is not measured (all measurements are on secondary characters) are computed using the previous results by now letting z_1, \dots, z_n denote the value of the character in n scored relatives and using

$$\mathbf{P}_\rho = \begin{pmatrix} 1 & \cdots & \rho_{z_1, z_n} \\ \vdots & \ddots & \vdots \\ \rho_{z_n, z_1} & \cdots & 1 \end{pmatrix} \quad (24.43a)$$

and

$$\mathbf{g}_\rho = \begin{pmatrix} \rho_{g_0, g} \\ \vdots \\ \rho_{g_0, g_n} \end{pmatrix} \quad (24.43b)$$

where the j th element in \mathbf{g} is the additive genetic correlation between the (unmeasured) individual of interest and its j th relative. The results from Equations 24.41a - 41d apply using these definitions.

Example 5. What is the index for predicting the breeding value in clutch size for a male given his mother's (z_1) and grandmother's (z_2) clutch size? From Table 3 in Chapter 6, $\rho_{g_0, g} = 1/2$ and $\rho_{g_0, g} = 1/4$. Assuming no epistasis and that shared environmental effects can be ignored, the phenotypic correlation between mother and grandmother is $h^2/2$. Thus

$$\mathbf{P}_\rho = \begin{pmatrix} 1 & h^2/2 \\ h^2/2 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{g}_\rho = \begin{pmatrix} 1/2 \\ 1/4 \end{pmatrix}$$

giving

$$\mathbf{b}_s = h^2 \mathbf{P}_\rho^{-1} \mathbf{g}_\rho = \left(\frac{h^2}{2(4-h^4)} \right) \cdot \begin{pmatrix} 4-h^2 \\ 1-h^2 \end{pmatrix} \quad \text{and} \quad \mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho = \frac{5+2h^2}{16}$$

$\sqrt{\mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho}$ ranges from a low of 0.56 when $h^2 \simeq 0$ to a high of 0.66 when $h^2 \simeq 1$, so that using the values from the mother and grandmother to construct

an index is about 60 percent as efficient as knowing an individual's phenotypic value.

Example 6. Consider the response under sib selection. Here n sibs are measured and based on the mean value of these individuals, the family is either accepted or rejected. If the family is accepted, other (unmeasured) sibs are used as parents to form the next generation. This is a model for selection when an individual must be sacrificed in order to reliably measure character value. Let r denote the additive-genetic correlation between sibs ($r = 1/4$ for half-sibs, $1/2$ for full-sibs) and t be the phenotypic correlation between sibs (the intraclass correlation coefficient). Under this design,

$$\mathbf{P}_\rho = \begin{pmatrix} 1 & t & \cdots & t \\ t & 1 & \cdots & t \\ \vdots & \vdots & \ddots & \vdots \\ t & t & \cdots & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{g}_\rho = r \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Using Equation 24.35 to obtain \mathbf{P}_ρ^{-1} gives (after a little algebra) the index for predicting the breeding value from an individual from the i th family as

$$I = \frac{h^2 r}{1 + (n-1)t} \sum_{j=1}^n (z_{ij} - \mu_z) = \frac{r n}{1 + (n-1)t} (\bar{z}_i - \mu_z)$$

where z_{ij} is the value of the j th sib in the i th family and $\bar{z}_i = n^{-1} \sum z_{ij}$ is the average value for this family. Likewise

$$\sqrt{\mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho} = \sqrt{\frac{n r^2}{1 + (n-1)t}} = r \sqrt{\frac{n}{1 + (n-1)t}}$$

which for large n approaches r/\sqrt{t} . Hence, the response to selecting individuals based entirely on the mean of their sibs is

$$\bar{i} \sigma_g h r \sqrt{\frac{n}{1 + (n-1)t}}$$

as obtained by Robertson (1955b) using a different approach. Note for large family size that both sib and between-family selection (Chapter 25 and next section) give essentially the same response. Sib and between-family selection differ only in that in the latter the individual is also measured. Thus for large family size, little is lost by not being able to measure an individual.

Within and between family selection. A particularly common set of relatives to consider are the family members of an individual and often selection is practiced using both individual and family values. For examine, Chapter 25 considers within-in family selection (individuals selected solely on their deviations from family mean) and between-family selection (whole families are either saved or culled depending solely on their mean). A number of other family-based selection schemes have been proposed, such as *sib selection* (Example 6) and *progeny testing* (using the mean value of an individual’s offspring). Turner and Young (1969) review applications of these in animal breeding, while Wricke and Weber (1986) review the special types of family selection possible in plants and other organisms with asexual reproduction and/or selfing.

Our concern here *combined selection*, which incorporates both within- and between-family selection by selecting on the index

$$I = b_1 \cdot (z - \bar{z}_f) + b_2 \cdot \bar{z}_f \tag{24.44a}$$

where \bar{z}_f is the individual’s family mean. Since b_1 weights the within-family deviation and b_2 weights the family mean, $(b_1, b_2) = (1, 0)$ corresponds to strict within-family selection, $(b_1, b_2) = (0, 1)$ to strict between-family selection, and $(b_1, b_2) = (1, 1)$ to individual selection. This index can also be expressed as

$$I = b_1 \cdot z + (b_2 - b_1) \cdot \bar{z}_f \tag{24.44b}$$

showing that within- and between-family selection can be simply related to an index combining individual and between-family selection.

Lush (1947) applied the Smith-Hazel index to obtain the optimal weighs for combined selection. To obtain his solution, consider the indirect index that optimizes the response in z given selection on the correlated characters $z_1 = z - \bar{z}_f$ (the within-family deviation) and $z_2 = \bar{z}_f$ (the family mean). To avoid separate expressions for half- and full-sib families, we express results in the notation of Chapter 25. Let t and r denote the phenotypic and additive genetic correlations (respectively) between sibs in an infinite population. These are related (Chapters 6, 25) by

$$t = rh^2 + c^2 \quad \text{with} \quad \frac{c^2}{\sigma_z^2} = \begin{cases} \sigma_{Ec(HS)}^2 & \text{for half-sibs} \\ \sigma_D^2/4 + \sigma_{Ec(FS)}^2 & \text{for full-sibs} \end{cases}$$

where $Ec(HS)$ and $Ec(FS)$ denote environmental effects common to half-sibs and full-sib families (respectively) and $r = 1/4$ for half sibs and $1/2$ for full sibs. In most cases half-sib families are formed by having a common father so that c^2 is expected to be negligible. Conversely, with full-sibs maternal effects can be quite important and hence c^2 considerable. Finally note that $c^2 < 1 - h^2$ so that c^2 is significant only when heritability is low. For a family of n sibs, the phenotypic

and additive genetic correlations have to be corrected slightly to account for finite population size, and we use the notation introduced in Chapter 25

$$t_n = t + \frac{1-t}{n} \quad \text{and} \quad r_n = r + \frac{1-r}{n}$$

To obtain the covariances required to construct the Smith-Hazel index, first note that $\sigma_{z_1, z_2} = 0$ as deviations from the mean and the mean itself are independent. Recalling Equations 25.16a and 16b, $\sigma_{z_2}^2 = t_n \sigma_z^2$ and $\sigma_{z_1}^2 = (1-t_n) \sigma_z^2$ giving the phenotypic correlation matrix as

$$\mathbf{P}_\rho = \begin{pmatrix} 1-t_n & 0 \\ 0 & t_n \end{pmatrix}$$

Similarly, Equations 25.19a and 19b give the vector of additive-genetic correlations of $(z_1, z_2)^T$ with z as

$$\mathbf{g}_\rho = \begin{pmatrix} 1-r_n \\ r_n \end{pmatrix}$$

Hence

$$\mathbf{P}_\rho^{-1} \mathbf{g}_\rho = \begin{pmatrix} \frac{1-r}{1-t} \\ \frac{1+n(1-r)}{1+n(1-t)} \end{pmatrix} = \begin{pmatrix} \frac{1-r}{1-t} \\ \frac{r_n}{t_n} \end{pmatrix}$$

giving the *Lush index* that optimally weights the within- and between-family effects as

$$I = (z - \bar{z}_f) + \left(\frac{r_n}{t_n} \right) \left(\frac{1-t}{1-r} \right) \bar{z}_f \quad (24.45a)$$

We have rescaled the index to emphasize the relative weighting of within-family deviation versus family mean. Figure 24.4 plots how these relative weights change as a function of t and n . If family size is infinite, within-family deviations and family means receive equal weight, and the Lush index reduces to individual selection (as $I = (z - \bar{z}_f) + \bar{z}_f = z$). For finite n , more weight is placed on within-family deviations when $t > r$ (phenotypic similarity between sibs exceeds their additive-genetic similarity) while family means receive more weight when $r > t$ (additive-genetic similarity exceeds phenotypic similarity). Significant family environmental effects are required for $t = rh^2 + c^2 > r$, so that within-family deviations receive more weight only if shared-family environmental effects are very important. The Lush index can be rearranged to assign weights to individual and family mean values,

$$I = z + \left(\frac{r-t}{(1-r)[1+n(1-t)]} \right) \cdot \bar{z}_f \quad (24.45b)$$

implying that family mean receives negative weight when $t > r$, as occurs when common environmental effects are very large. In such cases, much of the between-family differences are environmental rather than genetic and between-family differences are discounted in favor of within-family deviations.

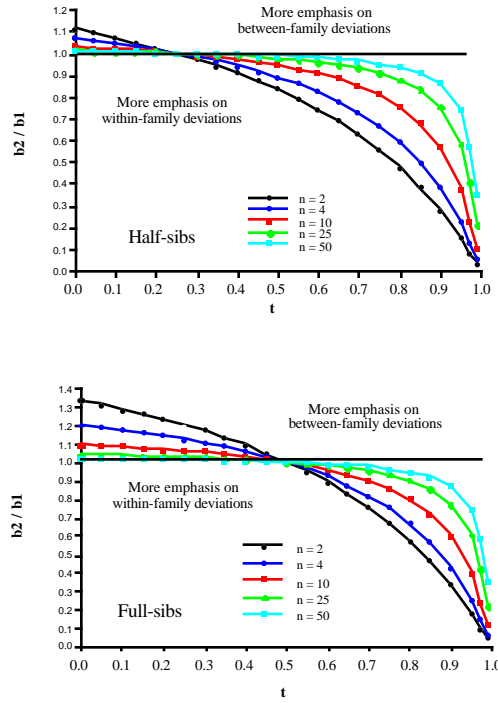


Figure 24.4. The relative weights under the Lush index of family means to within-family deviation (b_2/b_1) as a function of number of sibs n and correlation between sibs t . When $b_2/b_1 > 1$, more weight is placed on family mean (or equivalently, on between-family deviations), while more weight is placed on within-family deviations when $b_2/b_1 < 1$.

To obtain the expected response to selection on the Lush index, note first that

$$\mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho = 1 + \frac{(n-1)(t-r)^2}{(1-t)[1+t(n-1)]} \tag{24.46a}$$

giving (Equation 24.41d) the response in z as

$$\frac{R_z}{\bar{i}} = h \sigma_g \sqrt{1 + \frac{(n-1)(t-r)^2}{(1-t)[1+t(n-1)]}} \quad (24.46b)$$

More generally, consider the response to selection on the index $I = b_z(z - \bar{z}) + b_2 \cdot \bar{z}_f$ for arbitrary b_1 and b_2 . Taking the vector of characters as $\mathbf{z} = (z, z - \bar{z}_f, \bar{z}_f)^T$ and substituting $\mathbf{a} = (1, 0, 0)^T$ and $\mathbf{b} = (0, b_1, b_2)^T$ into Equation 15.5 gives the response in z as

$$\frac{R_z}{\bar{i}} = h \sigma_g \frac{b_1(1-r_n) + b_2 r_n}{\sqrt{b_1^2(1-t_n) + b_2^2 t_n}} \quad (24.47)$$

Figures 24.5 (half-sibs) and 24.6 (full-sibs) plots of responses of individual, strict within-family and strict between-family selection relative to the response under the Lush index. Note that r must be significantly different from t (additive-genetic similarity is much different from phenotypic similarity) for the index to be significantly superior to individual selection.

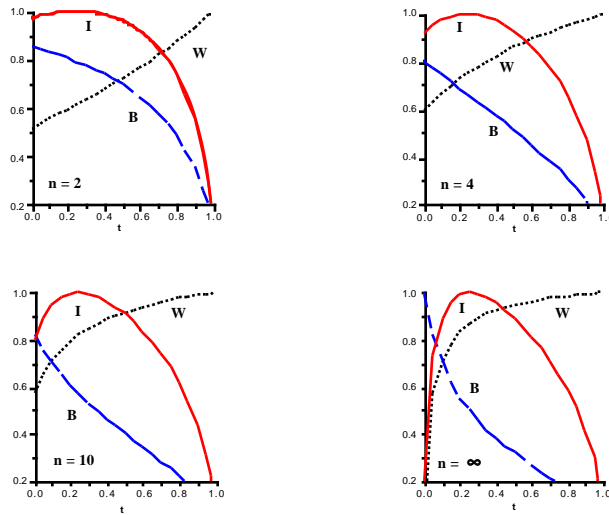


Figure 24.5. Expected single generation response in an infinite population of individual (I), strict within-family (W) and strict between-family (B) selection relative to that of the Lush index for half-sibs as a function of number of sibs n and correlation between sibs t . For half-sibs, it is generally expected that $t = h^2/4$ so that values of $t > 1/4$ occur only in highly usual situations.

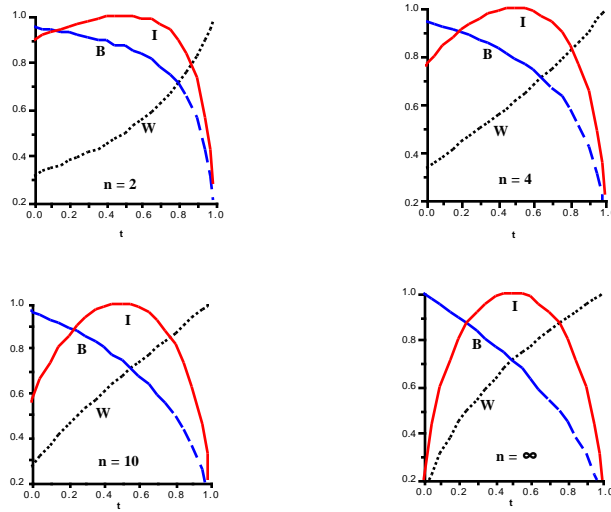


Figure 24.6. Expected single generation response in an infinite population of individual (I), strict within-family (W) and strict between-family (B) selection relative to that of the Lush index for full-sibs as a function of number of sibs n and correlation between sibs t .

One must be cautious of these comparisons of the relative efficiency of the Lush index as they are potentially misleading for several reasons. First, they assume selection intensities are the same in all comparisons, as one might (naively) expect if the same fraction of individuals is culled for each method. Finite population size results in overestimation the expected selection intensity. A second (and more subtle) source of overestimating the expected selection intensity is correlations between individuals, as occurs when multiple individuals from the same family are selected. Hill (1976, 1977c) examined this problem and gives tables of exact values and approximate expressions (which depend on n , t , and the number of families) for the expected selection intensities when individuals are correlated. When a small number of families is used, the selection intensity of the Lush index can be significantly below the value predicted by ignoring within-family correlations. Hence, proper comparisons must first correct for potential differences in the expected selection intensity.

A second concern is that these comparisons are correct for only a single generation of selection from an unselected base population. Selection generates gametic-phase disequilibrium (Chapters 25, 27) and increases inbreeding (Chapter 29), both of which have a larger effect on between-family selection. Gametic-phase disequilibrium reduces between-family additive genetic variance while

leaving within-family additive variance unchanged (Chapter 27), while selection entirely within a family results in less inbreeding (and hence less reduction in additive variance) than selection entirely between families (Chapter 29). As selection proceeds both these forces increase the importance of within-family effects relative to between-family effects, so that individual value becomes weighted more and family mean less. Wray and Hill (1989) note that while the relative efficiency of combined selection over individual selection may be greatly diminished by gametic-phase disequilibrium, the relative rankings of the methods still hold.

A final concern is that, as with any index, population parameters have to be correctly estimated else the index constructed from these estimates has incorrect weights and is less than optimal. Fortunately, for the Lush index only the intraclass correlation t must be estimated, and Sales and Hill (1976a) have shown that the efficiency of combined selection is quite robust to estimation errors in t (as initially suggested by Lush 1947).

Based on these concerns, it is not surprising that experimental verification of the advantage of the Lush index over individual or family selection is mixed. McBride and Robertson (1963) and Avalos and Hill (1981) found that combined selection gave a larger response than individual selection for abdominal bristles in *Drosophila melanogaster*. More conclusive results for selection on the same character were those of James (cited in Frankham 1982), who found that the observed increase in response under combined selection was $133 \pm 9.7\%$ and $111 \pm 7\%$ in two replicates, very consistent with the expected increase of 121%. Egg production in poultry was less conclusive, with Kinney et al. (1970) finding that individual selection gave a larger (but not significant) response than combined selection, while Garwood and Lowe (1981) found that combined selection gave a larger response (again not significant) than family selection. Larval and pupal weight in *Tribolium* showed similar mixed results, with Wilson (1971) finding that individual selection gave the largest response, while Campo and Tagarro (1977) did not find any significant differences (combined selection gave a larger response in a replicate with large family size, while individual selection showed the larger response in a replicate with small family size).

Finally, a more general combined index was considered by Osborne (1957b, c) which incorporates information from both full- and half-sib families. Osborne assumed the classic full-/ half-sib hierarchical design (Chapter 12) wherein a sire is mated to d dams, each of which has n sibs. Under this design each of the d families consists of n full sibs which are also half-sibs with respect to offspring from the other dams mated to the same sire. The resulting index to maximize response in z is constructed by considering three correlated characters: $z_1 = z - \bar{z}_{FS}$ (the deviation within full-sib families), $z_2 = \bar{z}_{FS} - \bar{z}_{HS}$ (the deviation between different full sib families from the same sire) and $z_3 = \bar{z}_{HS}$ where \bar{z}_{FS} is the mean of that individual's full-sib family (all offspring from the same dam) and \bar{z}_{HS} the half-sib mean of the individual (the mean of all offspring from that individual's sire). Denoting the intraclass correlation between half- and full-sibs by t_H and t_F ,

respectively, corresponding phenotypic correlation matrix is diagonal with

$$(nd) \cdot (\mathbf{P}_\rho)_{ii} = \begin{cases} d(n-1)(1-t_H-t_F) & \text{for } i = 1 \\ (d-1)[1-t_H+(n-1)t_F] & \text{for } i = 2 \\ 1-(n-1)t_F+(nd-1)t_H & \text{for } i = 3 \end{cases} \quad (24.48a)$$

and the vector of additive genetic correlations with z becomes

$$\mathbf{g}_\rho = (4nd)^{-1} \begin{pmatrix} 2d(n-1) \\ (d-1)(n+2) \\ 2+d+dn \end{pmatrix} \quad (24.48b)$$

and upon rescaling (to give full-sib family deviation weight one) the resulting index becomes

$$(z-\bar{z}_{FS}) + \frac{(2+n)(1-t_F-t_H)}{2(1+t_F(n-1)-t_H)} (\bar{z}_{FS}-\bar{z}_{HS}) + \frac{[2+d(1+n)][1-t_F-t_H]}{2[1+t_F(1-n)+t_H(dn-1)]} \bar{z}_{FS} \quad (24.49a)$$

A bit of algebra shows that $\mathbf{g}_\rho^T \mathbf{P}_\rho^{-1} \mathbf{g}_\rho$ equals

$$\frac{n-1}{4n(1-t_F-t_H)} + \frac{(d-1)(2+n)^2}{16dn(1+t_F(n-1)-t_H)} + \frac{(2+d+dn)^2}{16dn[1-t_F(n-1)+t_H(dn-1)]} \quad (24.49b)$$

Substituting into Equation 24.41d gives the expected response under this index. Osborne (1957b) presents graphs for the relative weights, but under the restrictive assumption of $t = r h^2$ (no dominance or common familial environmental effects). To construct the Osborne index only two parameters, t_F and t_H , must be estimated. Sales and Hill (1976a) show that although the index is more sensitive to poor estimates of t_H than of t_F , it (like the Lush index) is rather robust to errors in either estimated parameter.

Marker-assisted Selection

As was discussed extensively in Chapter 9, the recent ability to obtain polymorphic genetic markers spanning the genome at almost arbitrary small map distances has profound implications for quantitative genetics. Markers linked to QTLs (quantitative trait loci) can provide information on the genotype, and hence the breeding value, of an individual. Thus an individual's marker genotype can be viewed as a character (or set of characters) that can be used to improve selection response. While one might think that investigation of such *marker-assisted selection* (MAS) schemes is recent area of research, they were first considered by

Neimann-Sørensen and Robertson (1961) and Smith (1967). These authors developed selection indices for a single locus that directly influences the character of interest. Our treatment follows the more general approach Lande and Thompson (1990) who consider a selection index that allows for multiple marker loci linked to QTLs. Related approaches based on BLUP (best linear unbiased predictors) have been developed (Fernando and Grossman 1989, Fernando 1990, Cantet and Smith 1991, Goodard 1992), wherein marker information is used to construct the BLUP estimate for each individual's breeding value. Selection occurs by choosing those individuals with highest estimated values breeding values. These BLUP approaches are computationally very intense.

Assume a large number of markers have been scored and that at least some of these are in gametic-phase disequilibrium with QTLs underlying the trait of interest and hence provide some information about an individual's breeding value. How is this marker information translated into usable characters for index selection? To motivate the approach, consider the simplest case of a single diallelic locus (alleles Q and q) that is a QTL for the trait of interest. The contribution of this locus to the additive genetic value is $a + m\alpha$ for an individual with $m = 0, 1, 2$ copies of Q . The slope α of this regression of additive genetic value on number of copies of Q is the difference in average effects between alleles Q and q (Equations 4.9c, 4.10). This regression predicting additive genetic value as a function of the number of copies of a particular allele also holds more generally for a marker loci in gametic phase disequilibrium with a QTL (as opposed to the locus being the QTL itself). In this case, α decays to zero as the gametic phase disequilibrium between marker and QTL decays, so that a marker provides no information in the absence of disequilibrium with a QTL. Building on this single-locus regression, suppose now that a large number of diallelic marker loci are scored. We restrict analysis to each linkage group (operationally, each chromosome) under the assumption that disequilibrium between unlinked groups is very small. Suppose the linkage group being considered has n diallelic marker loci, the i th of which has alleles B_i and b_i . Letting $m_i = 0, 1, 2$ be the variable indicating the number of copies of allele B_i in the individual being examined, the average effects associated with the n markers are given by the slope coefficients of the least-squares regression of additive genetic value on the vector of marker genotypes $\mathbf{m}^T = (m_1, \dots, m_n)$,

$$g = a + \sum_{i=1}^n \alpha_i \cdot m_i = a + \boldsymbol{\alpha}^T \mathbf{m} \quad (24.50a)$$

Multiple regression is used instead of n separate univariate regressions for each marker loci because the genotypes of marker loci may be correlated with each other (marker loci in gametic phase disequilibrium) and/or a QTL may influence several markers at once (again generating correlations between loci). Multiple regression accounts for these correlated effects, with the resulting regression coefficients being the contribution from that locus with all others held constant (Chapter 7). Hence, $\alpha_i \cdot m_i$ is the estimated contribution towards additive genetic

value from the i th marker locus and we define m , the *marker score* of an individual, as the sum of the additive effects associated with these marker loci,

$$m = \sum_{i=1}^n \alpha_i \cdot m_i = \boldsymbol{\alpha}^T \mathbf{m} \tag{24.50b}$$

Marker score is the best linear predictor of additive genetic value given the vector of marker genotypes. Assuming no genotype-environment interactions, the observed phenotypic value z in places of the unobserved additive genetic value giving the estimate of $\boldsymbol{\alpha}$ from standard regression theory (Chapter 7) as

$$\hat{\boldsymbol{\alpha}} = \mathbf{S}_{\mathbf{m}}^{-1} \boldsymbol{\sigma}(z, \mathbf{m}) \tag{24.50c}$$

where $\mathbf{S}_{\mathbf{m}}$ is the estimated covariance matrix for \mathbf{m} (with $S_{m_{ij}} = \sigma(m_i, m_j)$) and $\boldsymbol{\sigma}(z, \mathbf{m})$ is the vector of estimated covariances between phenotype value and each single-locus marker genotype ($\sigma(z, \mathbf{m})_i = \sigma(z, m_i)$). The number of scored individuals must exceed the number of marker loci in order to uniquely estimate the marker effects for each loci. One practical point about constructing this regression is that only some of the markers are likely to be useful so that only a subset of the initially scored markers usually need to be followed. Care, however, is required in choosing the appropriate set of markers. The standard procedure is to choose markers by a stepwise multiple regression. Here the marker accounting for the most variation is used as the first variable in the regression. Next the marker accounting for the largest fraction of the remaining variance is examined and this marker is added to the regression if it explains a significant fraction of this variation. This procedure is repeated until the addition of a new marker does not account for a significant fraction of the remaining variance. The problem with this approach is that it produces biased overestimates of α_i . Lande and Thompson suggest a two-step approach where first a set of markers is chosen using stepwise regression on data from a previous generation and then a new regression computed using these chosen markers on the data from the current generation. The initial regression from the previous generate picks those markers that are potentially useful while the second regression using these markers on new data produces unbiased estimates.

Lande and Thompson show in many cases that the marker score is as informative as the entire vector of genotypic values, so that all marker information can be collapsed into the single variable m . Taking the vector of characters as $\mathbf{z} = (z, m)^T$, the vector of associated economic weights is $\mathbf{a} = (1, 0)^T$ as our concern is only response in the phenotypic value z . From Equation 24.29a, the vector of optimal weights is $\mathbf{P}^{-1} \mathbf{g}$, where \mathbf{g} is vector of additive genetic covariances associated with z . Let $\sigma_m^2 = \rho \sigma_g^2$ denote the variance in marker score, which accounts for ρ of the total variance in additive genetic value. Since m is a subset of genotypic values, $\sigma(m, g) = \sigma_m^2$, giving

$$\mathbf{g} = \begin{pmatrix} \sigma_g^2 \\ \sigma_m^2 \end{pmatrix} = \sigma_g^2 \cdot \begin{pmatrix} 1 \\ \rho \end{pmatrix} \tag{24.51a}$$

Since $\sigma(z, m) = \sigma(g + e, m) = \sigma(g, m) = \sigma_m^2 = \rho\sigma_g^2 = \rho h^2\sigma_z^2$, the associated phenotypic covariance matrix becomes

$$\mathbf{P} = \begin{pmatrix} \sigma_z^2 & \sigma_m^2 \\ \sigma_m^2 & \sigma_m^2 \end{pmatrix} = \sigma_z^2 \cdot \begin{pmatrix} 1 & h^2\rho \\ h^2\rho & h^2\rho \end{pmatrix} \quad (24.51b)$$

Applying Equation 24.29b, the Smith-Hazel index becomes

$$I_s = \mathbf{g}^T \mathbf{P}^{-1} \mathbf{z} = h^2 \left(\frac{1 - \rho}{1 - h^2\rho} \right) \cdot z + \left(\frac{1 - h^2}{1 - h^2\rho} \right) \cdot m \quad (24.52a)$$

which can be rescaled as

$$I_s = z + \left(\frac{1 - h^2}{h^2(1 - \rho)} \right) \cdot m \quad (24.52b)$$

Figure 24.7 plots the relative weights of marker score to phenotypic value as a function of h^2 and ρ . Marker score receives more weight when $\rho > 2 - 1/h^2$. Hence if marker score explains *any* additive genetic variance ($\rho > 0$), it receives more weight unless $h^2 > 1/2$.

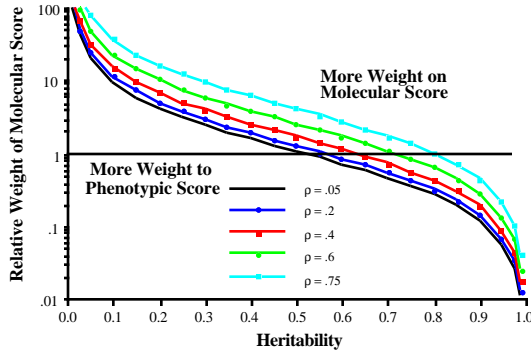


Figure 24.7. Ratio of the weights b_2/b_1 on molecular score m versus phenotypic value for the optimal index $I = b_1z + b_2m$ using marker loci and phenotypic value. Molecular score receives more weight when $\rho > 2 - 1/h^2$, where ρ is the fraction of additive genetic variance in z explained by m .

Noting that

$$\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g} = \sigma_g^2 h^2 \frac{1 + \rho/h^2 - 2\rho}{1 - h^2\rho} = \sigma_g^2 h^2 \cdot \left(1 + \frac{(1 - h^2)^2 \rho}{h^2(1 - h^2\rho)} \right) \quad (24.53)$$

Equation 24.29c gives the response under marker assisted selection as

$$\frac{R}{\bar{z}} = \sqrt{\mathbf{g}^T \mathbf{P}^{-1} \mathbf{g}} = \sigma_g h \cdot \sqrt{1 + \frac{(1 - h^2)^2 \rho}{h^2(1 - h^2 \rho)}} \quad (24.54)$$

This was obtained by Smith (1967) for a single marker locus and in the above more general form by Lande and Thompson (1990). Figure 24.8 plots the ratio of response under MAS to response under phenotypic selection solely on z . As expected, the response using marker information is more efficient than simple phenotypic selection (which has response $R/\bar{z} = \sigma_g h$). For a given heritability, as ρ approaches one the increase in response under MAS approaches its maximum value of $1/h$, so that the increase under MAS can be considerable if heritability is small. Conversely, as heritability increases towards one the response under MAS is not significantly different from simple phenotypic selection.

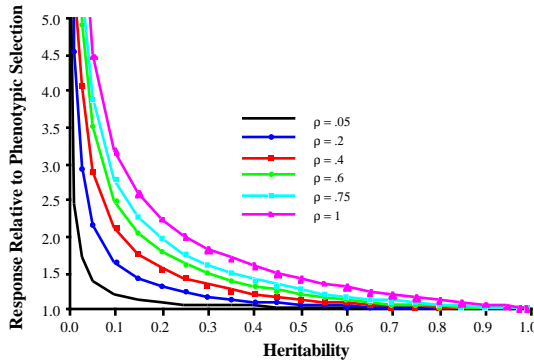


Figure 24.8. Ratio of the expected (single-generation) response of the optimal index for marker-assisted selection relative to selection solely on phenotypic value z .

The properties of marker-assisted selection have been examined in several simulation studies (Zhang and Smith 1992, 1993, Gimelfarb and Lande 1994). These confirmed MAS is more efficient than strict phenotypic selection, but that the increase in efficiency quickly declines as disequilibrium decays. Zhang and Smith (1992) noted that the disequilibrium between marker and QTLs decreases at a much quicker rate under marker-assisted selection than that predicted from random mating, which is not surprising since selection exploits the disequilibrium between marker and QTL. One candidate approach for generating new disequilibrium during MAS is to cross replicate lines after several generations of

selection, but Zhang and Smith (1992, 1993) show that this strategy is not efficient. Gimelfarb and Lande (1994) find that there is a significant advantage to reevaluating marker effects each generation, as this doubled the number of generations MAS was effective in their simulations. Somewhat surprising, they also found that increasing the number of markers does not necessarily increase the efficiency of response and that using too many markers can actually decrease the efficiency of MAS relative to MAS with fewer markers. This presumably arises because of estimation errors when a large number of markers are used. A related issue mentioned by Zhang and Smith (1993) is that MAS is only really efficient if QTLs that account for a significant fraction of the variance exist. As the distribution of QTL effects becomes more uniform and the number of QTLs increases, MAS becomes considerably less efficient and requires more markers (in which case the complications observed by Gimelfarb and Lande arise).

Indirect selection on marker score: applications to sex-limited traits. Marker score itself can be used as the sole basis of selection. From Equation 25.9b, the expected response in z to selection on m is

$$\frac{R_z}{\bar{i}} = \frac{\sigma_{m,z}}{\sigma_m} = \sigma_m^2 / \sigma_m = \sigma_m = \sigma_g \sqrt{\rho} \quad (24.55a)$$

giving an efficiency relative to selection on z of

$$\frac{\sigma_m}{h \cdot \sigma_g} = \sqrt{\frac{\rho}{h^2}} \quad (24.55b)$$

Thus as expected, marker selection gives a larger response than phenotypic selection when marker score accounts for more of the additive genetic variation than phenotypic value ($\rho > h^2$).

Marker-assisted selection can be used to enhance the response of a sex-limited character. Marker loci, by providing information on breeding value, can be used to select on the sex not expressing the character. Assume the sex-limited character is only expressed in females. Lande and Thompson (1990) consider the situation where the optimal index of phenotypic value and marker score is used to select females while marker scores are used to select males. Combining Equations 24.54 and 55a, the expected response is

$$R = h \sigma_g \left(\bar{i}_f \cdot \sqrt{1 + \frac{\rho(1-h^2)^2}{h^2(1-h^2\rho)}} + \bar{i}_m \cdot \sqrt{\frac{\rho}{h^2}} \right) \quad (24.56a)$$

where \bar{i}_f and \bar{i}_m is the selection intensity on females and male respectively. The expected response when selection occurs only on character value in females is $R = h \sigma_g \bar{i}_f$, showing that MAS increases response by

$$\sqrt{1 + \frac{\rho(1-h^2)^2}{h^2(1-h^2\rho)}} + \frac{\bar{i}_m}{\bar{i}_f} \cdot \sqrt{\frac{\rho}{h^2}} \quad (24.56b)$$

***Marker-assisted within- and between-family selection.** Lande and Thompson (1990) develop a modified Lush index for the optimal within- and between-family selection using both marker and phenotypic information. Letting \bar{z} and \bar{m} denote the mean phenotypic and marker scores for an individual's family, consider the index

$$I = b_1 \cdot (z - \bar{z}) + b_2 \cdot (m - \bar{m}) + b_3 \cdot \bar{z} + b_4 \cdot \bar{m}$$

so that b_1 and b_2 weight the within-family deviations in phenotypic value and molecular score, while b_3 and b_4 weight the family effects in these characters. Using arguments identical to those used for constructing the Lush index,

$$\mathbf{P} = \sigma_z^2 \cdot \begin{pmatrix} t_n & r_n h^2 \rho & 0 & 0 \\ r_n h^2 \rho & r_n h^2 \rho & 0 & 0 \\ 0 & 0 & (1 - t_n) & (1 - r_n) h^2 \rho \\ 0 & 0 & (1 - r_n) h^2 \rho & (1 - r_n) h^2 \rho \end{pmatrix} \quad (24.57a)$$

and

$$\mathbf{G} = \sigma_g^2 \cdot \begin{pmatrix} r_n & r_n \rho & 0 & 0 \\ r_n \rho & r_n \rho & 0 & 0 \\ 0 & 0 & (1 - r_n) & (1 - r_n) \rho \\ 0 & 0 & (1 - r_n) \rho & (1 - r_n) \rho \end{pmatrix} \quad (24.57b)$$

where (as above) $r_n = r + (1 - r)/n$ and $t_n = t + (1 - t)/n$. Since our interest is the response in $z = (z - \bar{z}) + \bar{z}$, the vector of economic weights is $\mathbf{a}^T = (1, 0, 1, 0)$. The resulting Smith-Hazel index is given by $I_s = (\mathbf{P}^{-1} \mathbf{G} \mathbf{a})^T \mathbf{z}$, or

$$\begin{aligned} & \left(\frac{r_n h^2 (1 - \rho)}{t_n - r_n h^2 \rho} \right) \cdot (z - \bar{z}) + \left(\frac{t_n - r_n h^2}{t_n - r_n h^2 \rho} \right) \cdot (m - \bar{m}) \\ & + \left(\frac{(1 - r) h^2 (1 - \rho)}{1 - t - (1 - \rho) h^2 \rho} \right) \cdot \bar{z} + \left(\frac{1 - t - (1 - r) h^2}{1 - t - (1 - \rho) h^2 \rho} \right) \cdot \bar{m} \end{aligned} \quad (24.58a)$$

Applying Equation 15.13, the response is $R_z/\bar{i} = \sqrt{\mathbf{a}^T \mathbf{G} \mathbf{b}_s}$, where

$$\begin{aligned} & \frac{\mathbf{a}^T \mathbf{G} \mathbf{b}_s}{h^2 \sigma_g^2} = \\ & \frac{\rho}{h^2} + (1 - \rho)^2 \left[\frac{r + (1 - r)/n}{t + (1 - t)/n - (r + (1 - r)/n) h^2 \rho} + \frac{(n - 1)(1 - r)^2}{n(1 - t - (1 - r) h^2 \rho)} \right] \end{aligned} \quad (24.58b)$$

Lande and Thompson present graphs for the increase in efficiency relative to the Lush index using only phenotypic values, which can also be obtained by comparing this with Equation 24.46b.

Marker considerations. Chapter 9 examined in detail many relevant issues of using markers to make inferences about QTLs, including how many individuals

must be scored to detect marker-QTL associations (also see Lande and Thompson 1990). Markers must be in gametic-phase disequilibrium with QTLs in order to provide information on an individual's breeding value, so that generating disequilibrium is a key issue in applying MAS. Three forces can generate disequilibrium: crossing populations with differing gametic frequencies (Chapter 5), genetic drift, and epistatic selection (Chapter 25). Epistatic selection is difficult to apply (a selection scheme use be developed that applies epistatic selection to markers and QTLs) so we focus on hybridization (population crosses) and small population sizes (genetic drift).

As discussed in Chapter 9, by far the most common and powerful approach for generating disequilibrium is hybridization. The amount of disequilibrium it generates is a function of both map distance and the differences in gamete frequencies between populations. Disequilibrium increases with population differences in gametic frequencies and with decreasing map distance. The optimal cross is two populations fixed for different marker-QTL linkage relationships (say *MMQQ* in population one and *mmqq* in population two), as occurs with crosses between inbred lines (Chapter 9). Recombination will remove any disequilibrium generated by hybridization, with the initial fraction of disequilibrium present after τ generations of random mating being $(1 - c)^\tau D_o \simeq e^{-c\tau} D_o$ (Equation 7.8). If the last hybridization event occurred τ generations ago, significant disequilibrium is expected only when the marker-QTL map distance $c < 1/\tau$. Finite population size also generates a small amount of disequilibrium. However, its effect is usually quite small requiring map distance $c < 1/4N_e$ for significant disequilibrium (Hill and Robertson 1968b). If effective population size N_e is even moderate disequilibrium is generated only for very tight linkage.

Jointly considering the effects of hybridization and drift, Lande and Thompson note that the map distance between marker and QTL must be less than the maximum of $1/\tau$ and $1/(4N_e)$ for significant disequilibrium to be present. Applying Equation 10.5, the number of randomly located markers required so that a fraction p of all QTLs are within this critical map distance $c^* = \max[\tau^{-1}, (4N_e)^{-1}]$ for a map of total length L is

$$\frac{\ln(1-p)}{\ln(1-2c^*/L)} = \frac{\ln(1-p)}{\ln(1 - \max[2/(\tau L), 2/(4N_e L)])} \simeq -\ln(1-p) \cdot \min[\tau L/2, 2N_e L] \quad (24.59a)$$

where we used $\ln(1-x) \simeq -x$ for $|x| \ll 1$ and $-\ln(1-p) \simeq .7, 2.3, 4.6$ for $p = .5, .9, .99$. This expression assumes a circular chromosome and underestimates the number of markers for linear chromosomes, with the underestimate becoming more severe as chromosome number increases. A related expression developed by Lande and Thompson (1990) is

$$\min[2\tau L, 8N_e L] + C \quad (24.59b)$$

where C is the haploid chromosome number. If selfing occurs rather than random

mating, the number of markers is greatly reduced to

$$4(1 - 1/2^t)L + C \quad (24.59c)$$

see Lande and Thompson (1990) and Lande (1992) for details.