Theoretical and Applied Genetics 46, 55-61 (1975)

Multi-Stage Index Selection

E.P. Cunningham

Animal Breeding & Genetics Department, The Agricultural Institute, Dunsinea, Castleknock, Co. Dublin (Ireland)

Summary. Selection index theory is extended to cover the case of selection in several stages. General algebra is given for adjusting in later stages for the effects of selection in earlier stages. In addition a method is developed for the incorporation of an index into an index. This simplifies the reuse of data from earlier stages of selection. A numerical example is used to illustrate the methods and to compare three single-stage and three two-stage selection procedures.

Introduction

When genes act additively, selection is the appropriate method for changing the genetic constitution of a population. The maximum gain from selection is obtained by using a selection index (Hazel, 1943; Henderson, 1963). The more complex the selection objective, and the more comprehensive the data to be considered, the more advantageous it becomes to use an index. Poor estimates of the current genetic structure of the population may make indexes inefficient (Heidhues and Henderson, 1962; Harris, 1964), but these conditions also undermine the basis for any alternative selection procedure. In general, therefore, selection is optimised by the use of an index.

In many modern plant and animal improvement programmes, selection is a continuous process, and several successive screenings may be applied in a single generation. For example, in the selection of dual-purpose bulls for use in artificial insemination in several European countries, the bulls are first selected using information about their parents; they are subsequently screened for their own growth potential in a performance test station, and finally selected on the basis of a comprehensive dairy progeny test of their daughters. The selection objective remains constant, but different information is used, and different selection intensity is applied at each stage. The final evaluation of a bull may be used as information in the selection of the next generation. There is thus a continuous cascade of information through time. With modern computing equipment, it should be possible to use all this information. In practice, each stage is most often treated as an entirely separate operation because of the theoretical and practical problem of dealing with selection at several levels. The purpose of this paper is to provide a theorectical framework and a data management algorithm which should make possible efficient multi-stage index selection.

Dickerson and Hazel (1944) give the method for dealing with the case of selection for a single trait in two stages. Its mathematical background is given by Cochran (1951). Jain and Amble (1962) extend this treatment to three stages. Papers by Cohen (1950), Finney (1956), Robson (1964), Young (1972) and Wang (1972) deal with aspects of the distributional properties of truncated populations and of their consequences in selection.

One Stage Selection

The information required in constructing a selection index can be specified in the following 4 vectors and 3 matrices.

$\underline{Y} = Y_1, \ldots, Y_m$	is a vector of additive genetic values for the m traits included in the aggregate genotype.
$\underline{v} = v_1, \ldots, v_m$	is a vector of constants, usually representing the relative economic values of the m traits in $\underline{Y}$.
$\underline{X} = X_1, \ldots, X_n$	is a vector of phenotypic measures for the n variables or sources of information be included in the index.
$\underline{b} = b_1, \ldots, b_n$	is a vector of weighting factors to be used in the index.
$\underline{P}$	is an $n \times n$ matrix of phenotypic covariances between the n variables in $\underline{X}$.
$\underline{G}$	is an $n \times n$ matrix of phenotypic covariances between the n variables in $\underline{X}$ and the m traits in $\underline{Y}$.

$\underline{C}$ is an $m \times m$ matrix of genotypic covariances between the m traits in $\underline{Y}$.

The aggregate genotype or breeding value is defined as

$$T = \underline{v'Y} = v_1Y_1 + v_2Y_2 + \ldots + v_mY_m.$$

Since T is not measurable, it cannot be selected for directly. Improvement in T is brought about by selection on an index or selection criterion:

$$I = \underline{b'X} = b_1X_1 + b_2X_2 + \ldots + b_nX_n.$$

The weighting factors in I are obtained by solving the index equations.

$$\underline{Pb} = \underline{Gv}$$

to give $\underline{b} = \underline{P^{-1}Gv}$

The variance of the index, the variance of the aggregate genotype and the covariance of index and aggregate genotype are

$$\sigma_I^2 = \underline{b'Pb} \qquad \sigma_T^2 = \underline{v'Cv} \qquad \sigma_{TI} = \underline{b'Pb}$$

The genetic change, in economic units, resulting from one round of selection is $\bar{i}\sigma_I$, where $\bar{i}$ is the selection differential achieved on a standardised distribution corresponding to the distribution of index values.

Two Stage Selection

Let the variables $\underline{X}_1 = X_1, \ldots, X_r$ be available for the first stage of selection. Let the additional variables $\underline{X}_2 = X_{r+1}, \ldots, X_n$ become available for the second stage.

Selection can then be done in one or two stages, and the data can be used in several ways:

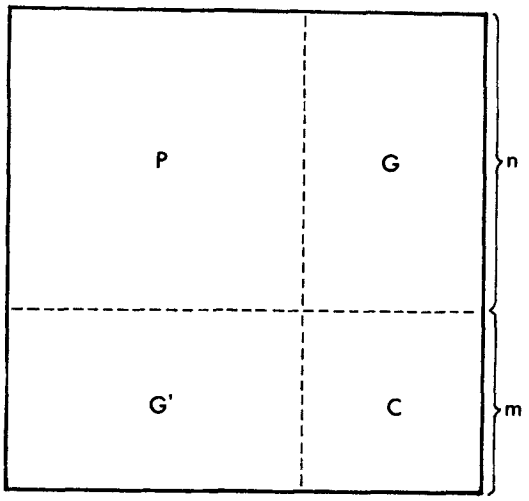
1. One-stage selection on $\underline{X}_1$ and $\underline{X}_2$,
2. One-stage selection on $\underline{X}_1$,
3. One-stage selection on I_1 and $\underline{X}_2$, where I_1 is the index that would have been used for option 2 above,
4. Two-stage selection, using $\underline{X}_1$ in the first stage and $\underline{X}_1$ and $\underline{X}_2$ in the second,
5. Two-stage selection, using $\underline{X}_1$ in the first stage and $\underline{X}_2$ in the second,
6. Two-stage selection, using $\underline{X}_1$ in the first stage and I_1 and $\underline{X}_2$ in the second.

Procedures 1 and 2 are straightforward applications of a selection index on variates $\underline{X}_1$, or $\underline{X}_1$ and $\underline{X}_2$. Procedures 4,5 and 6 require that all the variances and covariances linking $\underline{X}_1$ with $\underline{X}_2$ and $\underline{Y}$ be adjusted for the effects of selection on $\underline{X}_1$. Alternatives 3 and 6 require the incorporation of an index into an index. In the sections which follow, the general algebra for these two requirements is developed. A numerical example is then used to illustrate the methods and to compare the selection alternatives.

Incorporation of an index into an index

The variance-covariance matrix for the full set of variates and genotypic values can be represented by the following supermatrix.

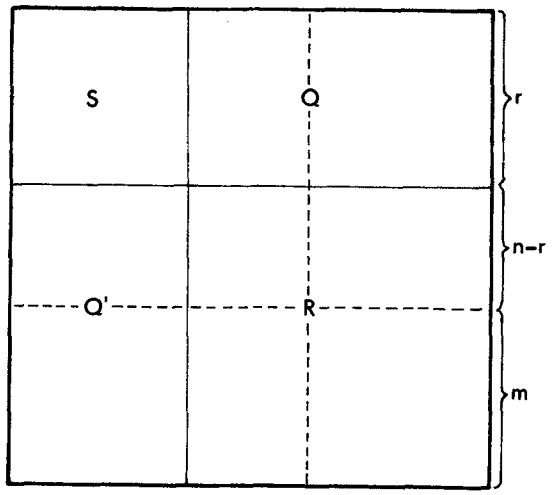
$\underline{M} =$
$E(\underline{X,Y})'(\underline{X,Y}) =$



(1)

If first stage selection is to take place on $\underline{X}_1 = \underline{X}_1, \ldots, X_r$ this matrix can be re-labelled as follows:

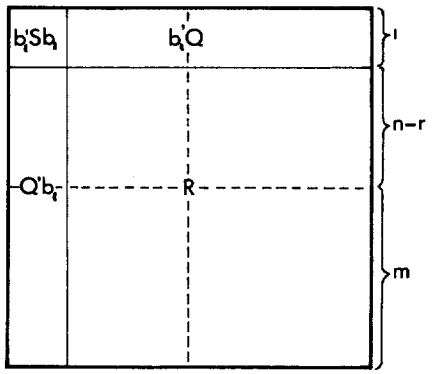
$\underline{M} =$
$E(\underline{X}_1, \underline{X}_2, \underline{Y})'(\underline{X}_1, \underline{X}_2, \underline{Y}) =$



(2)

If variates $\underline{X}_1$ are now replaced by their index $I_1 = \underline{b}_1' \underline{X}_1$, then a new reduced matrix can be constructed to contain the variances and covariances of the variates and genotypic values involved in stage 2.

$\underline{M}_r =$
$E(\underline{I}_1, \underline{X}_2, \underline{Y})' (\underline{I}_1, \underline{X}_2, \underline{Y}) =$



(3)

Note that the section ($\underline{R}$) of the original variance-covariance supermatrix involving $\underline{X}_2$ and $\underline{Y}$ is unaffected by collapsing $\underline{X}_1$ into an index. The first element of this new matrix is the variance of I_1, and the remainder of the first row and column contain the covariances of I_1 with the elements of $\underline{X}_2$ and $\underline{Y}$. If no prior selection has taken place on I_1, then this supermatrix can be repartitioned to give the input matrices needed to calculate an index on the variates $I_1, X_{r+1}, \ldots, X_n$, that is for selection procedure 3.

Adjustment for prior selection on I_1

If selection has taken place on I_1, then the variances and covariances in this matrix must be modified to take account of this. If there are any 3 normally distributed variates whose mutual covariances are known, and if truncation selection takes place on one of them, the covariances adjusted for the effects of selection can be obtained from a generalisation of formula (10) of Cochran (1951). Let the variates be z_i, z_j, and z_k, and their covariances be σ_{ij}, σ_{ik} and σ_{jk}. Assume truncation selection on z_i at a point t giving a standardised selection differential $\bar{i}$. These can be combined into a single selection parameter $s = \bar{i}(\bar{i} - t)$. A general expression for a covariance after selection is then

$$\sigma_{jk}^* = \sigma_{jk} - \sigma_{ij}\,\sigma_{ik}\,\sigma_{ii}^{-1}\,s \qquad (4)$$

This can be used to modify a whole matrix of covariances for the effects of selection on any one variate as follows. Let the ratio of the selection parameter, s, to the variance of the prior index, $\sigma^2_{I_1}$ be $w = s/\sigma^2_{I_1} = s/\underline{b}_1'\underline{S}\underline{b}_1$. In the case of two-stage selection with the second stage based on $\underline{I}_1$ and $\underline{X}_2$ (i.e. procedure 6) define a vector,

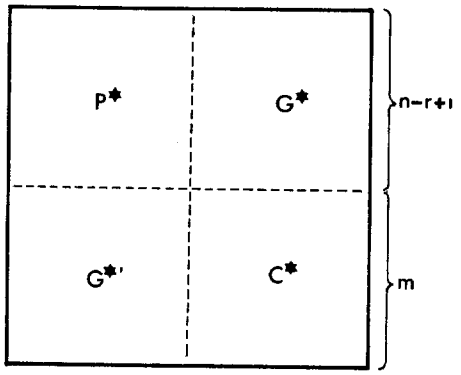
$$\underline{T} = (\underline{b}_1'\underline{S}\underline{b}_1, \underline{b}_1'\underline{Q}),$$

that is define it to be the first row of the matrix $\underline{M}_r$ in (3). The variances and covariances in $\underline{M}_r$ can then be adjusted for prior selection on I_1 by calculating

$$\underline{M}_r^* = \underline{M}_r - \underline{T}'\underline{T}\,w. \qquad (5)$$

This matrix can now be repartitioned to give the input matrices for a selection index based on $I_1, X_{r+1}, \ldots, X_n$ where prior selection has taken place on I_1. It becomes

$\underline{M}_r^* =$
$E^*(I_1, \underline{X}_2, \underline{Y})' (I_1, \underline{X}_2, \underline{Y}) =$



(6)

In the case of selection procedures 4 and 5, define a vector

$$\underline{T} = \underline{b}_1'(\underline{S}, \underline{Q})$$

that is, define it to be the product of $\underline{b}_1$ and the first r rows of the supermatrix $\underline{M}$ in (2). The variances and covariances in $\underline{M}$ can then be adjusted for prior selection on I_1 by calculating

$$\underline{M}^* = \underline{M} - \underline{T}'\underline{T}\,w. \qquad (7)$$

This supermatrix can now be repartitioned to give adjusted input matrices corresponding to those in (1).

Numerical Example

The data used in this example are taken from Hazel's (1943) classic paper on selection indexes. The selection objective or aggregate genotype in a swine selection

scheme consists of the three traits market weight (Y_1), market score (Y_2) and number of pigs born per litter (Y_3). Their relative economic weights are $\underline{v} = (^1/_3\ 1\ 2)$. The information available on which to base selection comprisss the five variates

X_1 = pig's own market weight
X_2 = pig's own market score
X_3 = productivity of dam
X_4 = average market weight of pig and littermates
X_5 = average market score of pig and littermates.

From the variances, heritabilities and correlations given by Hazel, it is possible to construct the $\underline{P}$, $\underline{G}$ and $\underline{C}$ matrices for the full system (1) as follows

$\underline{M} = E\,(\underline{X}, \underline{Y})\,'\,(\underline{X}, \underline{Y}) =$

1015.0596	93.5066	-7.4323	457.9949	41.2218	302.5893	13.5093	0.0
93.5066	22.8484	-3.7634	41.2218	8.2985	13.5093	2.2391	0.0
-7.4323	-3.7634	94.4784	-7.4323	-3.7634	0.0	0.0	7.6339
457.9949	41.2218	-7.4323	457.9949	41.2218	181.5536	8.1056	0.0
41.2218	8.2985	-3.7634	41.2218	8.2985	8.1056	1.3435	0.0
302.5893	13.5093	0.0	181.5536	8.1056	302.5893	13.5093	0.0
13.5093	2.2391	0.0	8.1056	1.3435	13.5093	2.2391	0.0
0.0	0.0	7.6339	0.0	0.0	0.0	0.0	15.2677

(8)

One-Stage Selection

If selection is to take place on all variates as a single operation (procedure 1), this system can be solved directly to give the index weighting factors and various measures of the effectiveness of the index. The weighting factors are

$\underline{b}' = (0.0976\ -\ 0.1654\ 0.1620\ 0.0867\ -\ 0.1892)$

The variance of the index is 17.6863 and its accuracy is 0.4087.

If information on the first three variates becomes available sooner than information on the remaining two, the vector of variates can then be divided into $\underline{X}_1 = (X_1 X_2 X_3)$ and $\underline{X}_2 = (X_4 X_5)$, and the supermatrix can be relabelled (2) accordingly:

$\underline{M} = E\,(\underline{X}_1, \underline{X}_2, \underline{Y})\,'\,(\underline{X}_1, \underline{X}_2, \underline{Y}) =$

If selection involves only $\underline{X}_1$ (procedure 2) the appropriate index can be found by simply ignoring $\underline{X}_2$ in solving this system. The weighting factors to use are

$\underline{b}_1 = \underline{S}^{-1}\underline{G}_1\,\underline{v}$

where $\underline{G}_1$ consists of the first three rows of $\underline{G}$. Their numerical values are $\underline{b}_1' = (0.1350\ -\ 0.2309\ 0.1630)$. The variance of this index is $\underline{b}_1'\,\underline{S}\,\underline{b}_1 = 16.3646$, and its accuracy or correlation with the aggregate genotype is 0.3932.

As specified in selection procedure 3, single-stage selection could be done using all five variates, but replacing the first three by their index $\underline{I}_1 = \underline{b}_1'\underline{X}_1$. Using formula (3), the variance - covariance matrix of all

1015.0596	93.5066	-7.4323	457.9949	41.2218	302.5893	13.5093	0.0
93.5066	22.8484	-3.7643	41.2218	8.2985	13.5093	2.2391	0.0
-7.4323	-3.7634	94.4784	-7.4323	-3.7634	0.0	0.0	7.6339
457.9949	41.2218	-7.4323	457.9949	41.2218	181.5536	8.1056	0.0
41.2218	8.2985	-3.7634	41.2218	8.2985	8.1056	1.3435	0.0
302.5893	13.5093	0.0	181.5536	8.1056	302.5893	13.5093	0.0
13.5093	2.2391	0.0	8.1056	1.3435	13.5093	2.2391	0.0
0.0	0.0	7.6339	0.0	0.0	0.0	0.0	15.2677

(9)

variates and traits involved becomes

$\underline{M}_r = E\,(I_1, \underline{X}_2, \underline{Y})\,'\,(I_1, \underline{X}_2, \underline{Y}) =$

16.3646	51.0997	3.0353	37.7433	1.3075	1.2437
51.0997	457.9949	41.2218	181.5536	8.1056	0.0
3.0353	41.2218	8.2985	8.1056	1.3435	0.0
37.7433	181.5536	8.1056	302.5893	13.5093	0.0
1.3075	8.1056	1.3435	13.5093	2.2391	0.0
1.2437	0.0	0.0	0.0	0.0	15.2677

(10)

If no selection has taken place on $\underline{X}_1$, this system can be solved to give an index in which the first three variates are replaced by their index I_1. Individuals would be selected using the criterion

$I_2 = 0.7886\, I_1 + 0.0793\, X_4 - 0.1951\, X_5.$

The accuracy of this index is 0.4070, and its variance is 17.5504.

Two-Stage Selection

In order to deal with the effect of stage 1 selection, it is necessary to specify the amount of selection involved. Assume that the upper 38 % of pigs on I_1 are selected. This gives a selection differential of $\bar{I} = 1.0$ and implies truncation at a point $t = 0.305$ on a standard normal distribution or $+ 0.305\, \sigma_{I_1}$ on the actual distribution of index values. The selection parameter is $s = \bar{I}(\bar{I} - t) = 0.695$ and $w = s/\sigma^2_{I_1} = 0.695/16.3646 = 0.0425$. Applying these values to the matrix (10) as described in (5) the matrix of variances and covariances adjusted for the effects of selection on I_1, is

$\underline{M}^*_r = E^*\,(I_1, \underline{X}_2, \underline{Y})\,'\,(I_1, \underline{X}_2, \underline{Y}) =$

This matrix can be used to calculate a second stage index based on variates I_1, X_4 and X_5 (i.e. procedure 6). It has the same weighting factors as I_2 above. In other words, the actual index is the same whether or not prior selection on I_1 has taken place. However because the variances and covariances of its constituents are reduced by selection on I_1, its variance is reduced to 6.1787. Its accuracy, or correlation with the aggregate genotype is 0.2557.

For comparison with this index, we could consider selecting solely on the information becoming available at the second stage (X_4 and X_5), but retaining the adjustments to variances and covariances which allow for prior selection on I_1 (procedure 5). This index is calculated by discarding the first row and column of matrix (11). It has a variance of 3.5661 and an accuracy of 0.1943.

To calculate a second stage index which uses all five variates individually (procedure 4), the original supermatrix $\underline{M}$ must first be adjusted for prior selection on I_1 by using formula 7. This adjusted supermatrix $\underline{M}^*$ can then be repartitioned to give the input matrices for the second stage index. The weighting factors which it produces are the same as where no prior selection has

4.9912	15.5854	0.9257	11.5117	0.3988	0.3793
15.5854	347.0989	34.6346	99.6432	5.2681	-2.6990
0.9257	34.6346	7.9032	3.2402	1.1750	-0.1603
11.5117	99.6432	3.2402	242.1881	11.4135	-1.9935
0.3988	5.2681	1.1750	11.4135	2.1666	-0.0690
0.3793	-2.6990	-0.1603	-1.9935	-0.0690	15.2021

(11)

taken place (option 1). However, the variance of the index and its accuracy are reduced to 6.3231 and 0.2308 respectively.

In order to compare the effects of these different selection procedures, it is necessary that they all have the same final intensity of selection. Let this be 6 %. In the case of one-stage selection, this gives a selection differential of two standard deviations. Equivalent selection in the case of a two-stage procedure can be achieved by taking the best 38 % on the first stage and the best 16 % (i.e. 6/38) of these remaining on the second stage. This gives a selection differential of one standard deviation for stage one, and 1.52 standard deviations for stage two.

The net effect of selection on any index, in economic units, is $\bar{i}\sigma_I$ where $\bar{i}$ is the standardised selection differential and σ_I is the standard deviation of the index. The relative effectiveness of the six options considered can therefore be given as follows (Table 1).

Table 1. Relative effectiveness of the six selection procedures

	Selection Procedure	$\bar{i}$	σ_I	Gain from selection Absolute	Relative
1.	One stage, variates $X_1X_2X_3X_4X_5$	2.0	4.2055	8.4110	100.0
2.	One stage, variates $X_1X_2X_3$	2.0	4.0453	8.0906	96.2
3.	One stage, variates $I_1X_4X_5$	2.0	4.1893	8.3786	99.6
4.	Two stage, variates $X_1X_2X_3$ in stage 1	1.0	4.0453	7.8675	93.5
	variates $X_1X_2X_3X_4X_5$ in stage 2	1.52	2.5146		
5.	Two stage, variates $X_1X_2X_3$ in stage 1	1.0	4.0453	6.9157	82.2
	variates X_4X_5 in stage 2	1.52	1.8884		
6.	Two stage, variates $X_1X_2X_3$ in stage 1	1.0	4.0453	7.8236	93.0
	variates $I_1X_4X_5$ in stage 2	1.52	2.4857		

Discussion

The extension of the method to more than two stages of selection is straightforward. The supermatrices $\underline{M}^*$ or $\underline{M}^*_r$ adjusted for prior selection on I_1 could be regarded as the starting point (1) for a two stage selection, thus giving three stages in all. This can be repeated as often as required. The main requirement is that the full variance-covariance matrix of all variates to be used and traits to be selected for must be available at the beginning.

One convenient result is that the actual index weighting factors are not affected by prior selection. This of course is the same thing as saying that a regression coefficient is unaffected by selection on the independent variate. In practice this simplifies the application of multi-stage index selection, since new indexes need not be calculated at different stages and for different intensities of prior selection.

The calculation of a selection index does not depend on the form of multivariate distribution linking the variates and traits involved. However, the calculation of the effects of selection on the index is distribution dependent. The methods given in this paper assume an initial multivariate normal distribution. The results up to and including the index to be used in the second stage should therefore be exact. The estimated gain from second stage selection will tend to be overestimated, since the selection differential used relates to a normal distribution, whereas the distribution of index values can be expected to depart from normality. At the third stage, both the index and its estimated effects will be biased because the input matrices are adjusted on the assumption that they were multivariate normal at stage 2. There appears to be no exact solution for this problem, even for the case where the vectors of variates $\underline{X}$ and of traits $\underline{Y}$ can each be regarded as a single variable. Jain and Amble (1962) discuss alternative methods of dealing with it, while Young (1972) gives a method of numerically evaluating the mean and variance of selected populations for up to four stages of selection.

The gradual deformation of the initial normal distribution can be quite severe if selection is intense, if one can judge from the effects in the bivariate case (Cochran, 1951). It may be that with an array of variates in the selection criterion and an array of traits in the selection objective, that the normality of their multivariate distribution is better protected than in the case of single variate selection for a single trait objective. It may also be that if two stages are separated by a generation, that genetic segregation and recombination have the effect of recreating normality in each generation. At any rate, this problem of the decay of normality under repeated selection is one which this paper does not attempt to resolve.

It is encouraging that the more convenient procedure 6 is apparently almost as good as the full two stage procedure 4. In practice, it would often by very inconvenient to be required to reassemble the original data used in the first stage of selection for reuse in the second stage. However, if much the same result can be achieved merely by using the index values calculated for the first-stage as input variates for the second-stage, then the practical possibilities of efficient multi-stage index selection are greatly increased.

Acknowledgements

I wish to thank Prof. A.L. Rae, Mr.J. Connolly, Mr.P. Brascamp, Prof. D.J. Finney and Prof. Alan Robertson for their useful comments.

Literature

Cochran, W.G.: Improvement by means of selection. Proc. 2nd Berkeley Symp. on Math. , Stat. and Probability, ed. Neyman, J. 449-470 (1951)

Cohen, J.A.C.: Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. Annals Math. Stat. 21 , 557-569 (1950)

Dickerson, G.E.; Hazel, L.N.: Effectiveness of selection on progeny performance as a supplement to earlier culling in livestock. J. Agric. Res. 69, 459-475 (1944)

Finney, D.J.: The consequences of selection for a variate subject to errors of measurement. Rev. Inst. Intern. Statistique 24, 1-10 (1956)

Harris, D.L.: Expected and precided progress from index selection involving estimates of population parameters. Biometrics 20, 46-72 (1964)

Hazel, L.N.: The genetic basis for constructing selection indexes. Genetics 28, 476-490 (1943)

Heidhues, T.H.; Henderson, C.R.: Beitrag zum Pro-Problem des Basisindex. Z.f. Tierez. 77, 291-311 (1962)

Heidhues, T.H.; Henderson, C.R.: Selection index and expected genetic advance. In: Statistical genetics and plant breeding, ed. Hanson, W.D. and Robinson, H.F. Natl. Acad. Sci. Nat Res. Council Publ. 982 111-163 (1963)

Jain, J.P.; Amble, V.N.: Improvement through selection at successive stages. J.Ind. Soc. Agr. Stat. 8, 88-109 (1962)

Robson, D.S.: Note on repeated selection in the normal case. Report No. BU-171-M. Cornwall University (1964)

Wang, Y. Ying: Selection problems under multivariate normal distribution. Biometrics 28 , 223-233 (1972)

Young, J.C.: The moments of a distribution after repeated selection with error. J. Am. Stat. Assoc. 6, 206-210 (1972)

Received September 25, 1974
Communicated by A. Robertson

Prof. E.P. Cunningham
Animal Breeding and Genetics Dpt.
The Agricultural Institute
Dunsinea Research Centre
Castleknock, Co. Dublin (Ireland)