DESIGN EFFICIENCIES WITH PLANNED AND UNPLANNED UNBALANCE FOR ESTIMATING HERITABILITY IN FORESTRY

Barbara G. McCutchan[1/]

Abstract.--Both balanced and unbalanced data can be analyzed for variance component estimation with Modified Maximum Likelihood estimates in a unified approach. Design efficiencies are evaluated for the estimation of heritability using this methodology, assuming knowledge of the variance components. Rules for obtaining efficient randomized block designs are established. The effect of number of blocks, plot size, number of families, variance on family size and total number of observations on design efficiency is examined across the range of heritability and under 100%, 90%, 80% and 60% survival.

Additional keywords: Modified Maximum Likelihood, Design allocation rules.

INTRODUCTION

One of the problems that the experimenter faces in forestry designs for the estimation of means and also variance components is the use of large blocks in balanced experimental designs. Large blocks necessitate employment of either a restricted set of environments where small plot variances can be found or the inclusion of block-type variation among plots within blocks. In this latter situation, the error in estimating family means is increased, effectively decreasing heritability.

Anderson (1975, 1981) and others suggest and evaluate intentionally unbalanced designs for the estimation of variance components. These planned unbalanced designs allow for the redistribution of the degrees of freedom to the variance components of interest. Various unbalanced two-way designs, useful for forest genetics trials of half-sib families with small randomized blocks, are also possible to design (McCutchan 1985).

A complicating factor in most forestry experiments, and one which makes design evaluation difficult, is that some level of unplanned loss occurs in a genetic trial subsequent to its establishment. Roughly 10% loss occurs in loblolly pine (*Pinus* taeda L.) genetic trials after one year of field growth, with up to 30% loss occurring by age 10, depending upon the incidence of fusiform rust (*Cronartium quercuum* f. sp. *fusiforme*) (R. J. Weir, pers. comm.). The efficiency of a design for the estimation of particular variance components or functions thereof, under states of loss, is of consequential interest.

[1/]Quantitative Geneticist, Westvaco Forest Research, Summerville, SC. This research was conducted while the author was a Graduate Research Assistant at North Carolina State University, Raleigh, NC. The author graciously acknowledges Drs. Gene Namkoong and Francis Giesbrecht for their guidance with this research, the Statistics Department at North Carolina State University, Raleigh, for providing the necessary computing funds, Ronnie Hise for his drawing of the graphics, and the North Carolina State Hardwood Research Cooperative, Raleigh, NC, for releasing the 1981 sycamore data.

The analysis of unbalanced data--either planned or unplanned unbalance--is therefore important for two reasons: (1) after an experiment is completed, an efficient estimator is needed and (2) before an experiment is conducted, designs need to be analyzed for the possible efficiency with which parameter estimates will ultimately be made (Namkoong 1981). Unfortunately, the most commonly used analytical procedure, namely Henderson's Method 3 (1953), is ambiguous as to which sum of squares is most appropriate; such estimators retain only their unbiased property with unbalanced data. The maximum likelihood type estimators have been known theoretically, but have not been available practically. Giesbrecht (1983) has written an efficient algorithm by which Modified Maximum Likelihood (MML), Maximum Likelihood (ML) and MInimum Norm Quadratic Unbiased Estimates (MINQUE) can be computed. The MML approach is used for the remainder of this paper. The MML estimates, for which normality is assumed, are chosen because of their desirable properties regardless of the state of balance in the data. The estimates maximize the likelihood, use the same information as the full ML estimates do and account, in some sense, for the estimation of fixed effects; with balanced data, MML estimates are also those obtained through the analysis of variance (AOV), which is not true for ML estimates. The MML estimates are obtained by iterating the MINQUE. The MML method is a unified approach to the estimation of variance components and/or for comparing design efficiency.

The objective of this paper is to compare design efficiencies of planned balanced and unbalanced designs for the estimation of heritability (h'). The unbalanced designs allow for the inclusion of a large number of families in relatively small blocks. The variance of the estimate of h^2 ($var(h^2)$) from each design is compared to other designs across the range of h^2 and with 10%, 20% and 40% random loss of individuals. Design efficiency is examined over the range of h^2 to indicate the quality of the design at any level of realized h^2 or for multiple traits which may have different h^2 in the same experiment. The design structure studied is a randomized block design on one location; the treatments are unrelated half-sib families using either single-tree or two-tree contiguous plots. The variance components are assumed known which enables calculation of the variances of the variance components. An overview of the results from McCutchan et al. (submitted) and McCutchan (1985) is presented.

METHODOLOGY

The notation and the computational methodology for Modified Maximum Likelihood follow Giesbrecht's (1983). His procedure for variance component estimation is written as a temporary Statistical Analysis System (SAS') program entitled Procedure MIXMOD.

The statistical model for each design considered is:

$$\underset{nx1}{Y} = \mu \underset{nx1}{1} + \underset{nxb}{U_B} \underset{bx1}{e_B} + \underset{nxf}{U_F} \underset{fx1}{e_F} + \underset{nxs}{U_P} \underset{sx1}{e_P} + \underset{nx1}{e_W}$$

where Y is the column vector of n observations; p is the overall mean; U_B, OF and U_P are design matrices pertaining to block, family and plot effects, respectively, with all elements equal to zero or one (where there are b blocks, f families and s filled combinations of the families and blocks); for

single-tree plots, U_P is the identity matrix of size n; e_B, e_F, e_P and e_W are independent column vectors of independent random variables, each with zero mean and variance-covariance matrix $I_b\sigma_B^2$, $I_f\sigma_F^2$, $I_s\sigma_P^2$ and $I_n\sigma_W^2$, respectively.

The variance-covariance matrix for the vector of observations (Y) is:

$$V(Y) = U_B U_B'\sigma_B^2 + U_F U_F'\sigma_F^2 + U_P U_P'\sigma_P^2 + I_n\sigma_W^2 ,$$

where σ_B^2, σ_F^2, σ_P^2 and σ_W^2 are the variance components due to the block, family, plot and within-plot effects, respectively. Letting $V_i = U_i U_i'$ and for convenience $V_W = I_n$, V(Y), based on the parameters, can be rewritten as:

$$V_{\sigma^2} = V(Y) = V_B\sigma_B^2 + V_F\sigma_F^2 + V_P\sigma_P^2 + V_W\sigma_W^2 . \qquad (1)$$

It is assumed that the f unrelated families chosen are a random sample of the reference population, that those trees planted are all to be assessed save for those lost and that blocks of different sizes are placed on different parcels of land such that for a given set of variance components $(\sigma_B^2\ \sigma_F^2\ \sigma_P^2\ \sigma_W^2)'$ different sizes of blocks and plots can be compared.

If the variance components (σ_i^2) were known, then the variance-covariance matrix for the resulting MML estimates of the components $(\sigma_B^2\ \sigma_F^2\ \sigma_P^2\ \sigma_W^2)'$, which would then be MInimum Variance Quadratic Unbiased Estimates, would be:

$$2\{tr(Q_{\sigma^2}V_i Q_{\sigma^2}V_j)\}^{-1} \qquad i,j = B,F,P,W \qquad (2)$$

where $Q_{\sigma^2} = V_{\sigma^2}^{-1} - V_{\sigma^2}^{-1}\,1\,(1'V_{\sigma^2}^{-1}1)^{-1}1'V_{\sigma^2}^{-1}$, and V_{σ^2} is defined in Eq. 1.

The dispersion matrix (Eq. 2) of the variance components is a function of the variance components themselves plus the design matrices (U_i). It is therefore possible to calculate this dispersion matrix for a given set of true variance components and a design. The observational values (Y) are not needed to calculate the dispersion matrix for the variance components. Heritability is calculated based on the experimental structure as:

$$h^2 = 4\sigma_F^2/(\sigma_F^2 + \sigma_P^2 + \sigma_W^2) = X/Y. \qquad (3)$$

Since variance component estimation is based on the experimental design in only one environment, any environment by additive genetic interaction variance is confounded with the estimates of the additive variance σ_A^2. Such an estimate is appropriate for inferences only on this site type. The variance of

the estimate of h^2 can be approximated by the variance of a ratio using a Taylor's series expansion:

$$\text{var}(\hat{h}^2) \approx (1/Y)^2\text{var}(\hat{X}) - 2(1/Y)(X/Y^2)\text{cov}(\hat{X},\hat{Y}) + (X/Y^2)^2\text{var}(\hat{Y}). \quad (4)$$

The computation of $\text{var}(\hat{h}^2)$ is based on the calculation of $\text{var}(\hat{X})$, $\text{cov}(\hat{X},\hat{Y})$ and $\text{var}(\hat{Y})$ from the dispersion matrix (Eq. 2). The values of X and Y are calculated from the set of variance component parameters. The approximation of $\text{var}(\hat{h}^2)$ relies on large sample theory.

The $\text{var}(\hat{h}^2)$ is calculated for each of the designs in Figure 1 for two types of variance component sets. Only the type $\sigma_F^2 = \sigma_B^2$ is reported here, where $\sigma_B^2 = 2$, $\sigma_W^2 = 1$ and $\sigma_F^2 = \sigma_P^2$ take on values of .5, .1, .05 and .005 for h^2 of 1.0, .33, .18 and .02, respectively. The actual assignment of the families to the blocks is not given, as it was found to be inconsequential in terms of design efficiency (McCutchan et al. submitted). In each design the effect of the random loss of 10%, 20% and 40% of the individuals on design efficiency is examined. The $\text{var}(\hat{h}^2)$ reported for cases of loss is actually an average of two independent samplings of individual loss.

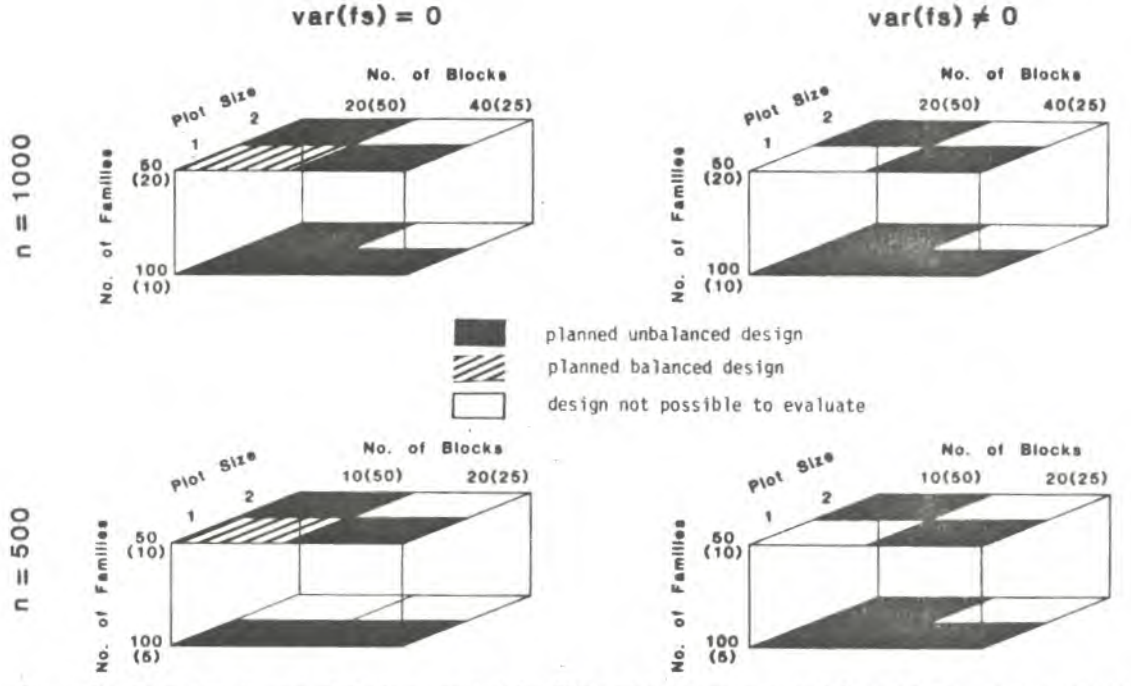

Figure 1. Schematic diagram of designs studied for 1000 and 500 observations (n), and for equal (var(fs)=0) and variable (var(fs)≠0) family sizes. The var(fs) is proportional to the mean family size and to n. The number in parentheses indicates the size of the effect; in the case of variable family size, the number is the mean family size.

The effects of block size, plot size, family size, variance of family size and total experimental size on design efficiency are each assessed at four levels of survival across the range of h^2. To examine any one effect, the var($\hat{h}^2$) from each of two designs, which differ only in their levels of this effect, are compared in a ratio. If the ratio is one, then one level of the effect is just as efficient as the other level. If the ratio is greater than one, then the design whose var($\hat{h}^2$) is in the numerator of the efficiency ratio is less efficient. In assessing the efficiency ratios for each effect, the buffering to loss is discussed as is the comparison of var($\hat{h}^2$) to a particular criterion (Namkoong and Roberds 1974). The criterion is established on a CV = 50% = $(\sqrt{\text{var}(\hat{h}^2)}\ /h^2) \times 100\%$ for $h^2 \geq .2$ and on std($\hat{h}^2$) = .10 = $\sqrt{\text{var}(\hat{h}^2)}$ for $h^2 \leq .2$. Ideally, a design is sought whose var($\hat{h}^2$) profile falls beneath that of the standard across the range of h^2.

RESULTS AND DISCUSSION

All design efficiency ratios are computed based on a given set of variance components (σ^2_B σ^2_F σ^2_P σ^2_W), regardless of block, plot or family size.

The effect of block size on design efficiency is illustrated in Table 1 for the 100-family single-tree plot (STP) design with 1000 observations. The ratio of the var(h^2) from the 20-block block design is less than that from the 40-block design across the range of h^2. The larger block design is uniformly more efficient than a design with more smaller blocks, given the same set of variance components.

The effect of random loss on such a comparison is shown in Table 2 for h^2 = .33. The ratio decreases with loss indicating that the larger block design is better buffered to loss. This is true also across the range of h^2. The specifics for other comparisons of the block effect are given by McCutchan (1985).

Table 1.--Design efficiency ratios as affected by number of blocks for 100% survival, 100 families, STP and n=1000

h^2	Number of Blocks	
	40	20
	var($\hat{h}^2$)	$E_{b=20}$ 1/
1.00	.0219	.982
.33	.0096	.977
.18	.0068	.975
.02	.0041	.973

Table 2.--Design efficiency ratios as affected by number of blocks for h^2 = .33, 100 families, STP and n=1000

Survival (%)	Number of Blocks	
	40	20
	var($\hat{h}^2$)	$E_{b=20}$
100	.0096	.977
90	.0109	.974
80	.0128	.970
60	.0195	.958

1/ $E_{b=20}$ is the ratio of the var($\hat{h}^2$) from the 20-block design divided by that from the 40-block design.

The initial motivation for examining the usefulness of smaller blocks was the observation that smaller homogeneous sites are more frequent than larger homogeneous sites. In these comparisons designs have been examined

for the same set of variance components over the range of h^2 and a range of random loss, regardless of block and plot size. The designs with larger blocks are 2% to 3% more efficient than those with 25-tree plots, and are better buffered to loss, given the same set of variance components. The practical application of these results includes consideration of the frequency at which these larger sites can be found. For a fixed n, fewer larger sites would be required than small sites; whether b large blocks could be found for a given site type, of course, depends upon the site. Considering the use of 20 blocks, the results show that by using designs with blocks half the size, a 2% to 3% loss in efficiency is incurred. (The loss in efficiency indicates the increase in var(h^2) in having used 40 blocks versus 20 blocks). This cost in efficiency has to be balanced against the cost of obtaining and maintaining half as many blocks, each of twice the size. This latter cost may include not only difficulties in locating such blocks, but also bias in representing planting sites.

The effect of plot size on design efficiency is illustrated in Table 3 with the comparison of a single-tree plot (STP) design to a two-tree plot (TIP) design given 100 families, 20 blocks and n=1000. The STP design is more efficient than that with TTP across the range of h (Table 3), with the advantage in efficiency at 100% survival decreasing with h2. The STP design remains more efficient with the imposition of random 10%, 20% and 40% loss (Table 4).

Table 3.--Design efficiency ratios as affected by plot size for 100% survival, 100 families, 20 blocks and n=1000

h^2	Plot Size	
	STP	TTP
	var($\hat{h}^2$)	E_{TTP} [1/]
1.00	.0215	1.2351
.33	.0094	1.1815
.18	.0066	1.1676
.02	.0039	1.1570

Table 4.--Design efficiency ratios as affected by plot size for h^2 = .33, 100 families, 20 blocks and n=1000

Survival (%)	Plot Size	
	STP	TTP
	var($\hat{h}^2$)	E_{TTP}
100	.0094	1.1815
90	.0107	1.1768
80	.0124	1.1685
60	.0187	1.1675

1/ Epp is the ratio of the var(h^2) from the TTP design divided by that from the STP design.

The premise of using a TTP design is to protect the data set against loss of plots. An AOV can be used for balanced data on a plot mean basis. Loss of plots is not a computational or interpretative obstacle with the MML methodology. There is a statistical cost to using TTP, as observed here, which even at that fails to insure plot survival.

The large number of small family design is more efficient for 100% survival and high heritabilities than the small number of large family design (Table 5, E_f=50). This result is reversed for low heritabilities, where the larger family design is more efficient. The effect of random loss on the design efficiency is given in Table 6by $E_{f=50}$ based on 10%, 20% and 40% random loss. At each h^2 given, the var(h^2) from the 50-family design increases

less than that from the 100-family design. The buffering capacity of the design to loss is greatly influenced with these size differences in families, there being roughly twice as much buffering capacity at $h^2 = 1.0$ for the 50-family design compared to the 100-family design, with this difference decreasing with h^2. This greater buffering capacity with 50-family designs is reflected in a decreasing $E_{f=50}$ with loss. The 50-family design, with this large difference in buffering capacity, becomes more efficient with 10% loss at h = .33 in contrast to the block or plot effects. In neither of these designs are families lost through random loss of individuals.

Table 5.--Design efficiency ratios as affected by number of families for 100% survival, 40 blocks, STP and n=1000

h^2	Number of Families	
	100	50
	$var(\hat{h}^2)$	$E_{f=50}$ 1/
1.00	.0219	1.476
.33	.0096	1.026
.18	.0068	.824
.02	.0041	.517

Table 6.--Design efficiency ratios as affected by number of families for h^2 = .33, 40 blocks, STP and n=1000

Survival (%)	Number of Families	
	100	50
	$var(\hat{h}^2)$	$E_{f=50}$
100	.0096	1.026
90	.0109	.982
80	.0128	.935
60	.0195	.830

1/ $E_{f=50}$ is the ratio of the $var(\hat{h}^2)$ from the 50-family design divided by that from the 100-family design.

The implications for design recommendations are that STP and large blocks provide low $var(F^2)$ across the range of h^2, but that the family size that should be employed depends on the heritabilities of interest. The 100-family design is more efficient across a large portion of the range of h^2. If design allocations included only balanced designs, this efficient 100-family, 20-block design would not be a viable alternative. If all the traits of interest have low heritabilities, for example, less than .2, then a 100-family design would not be the most efficient design to use. The 50-family design would be more efficient in this range and have greater buffering to loss.

The effect of variable family size in contrast to equal family size on design efficiency is illustrated in Table 7 for 100 families of average size 10. The equal family size design is more efficient at $h^2 > .33$ than the variable family size design at 100% survival. The variable family size design is more efficient below this level of h^2, increasingly so as h^2 decreases. These results confirm the suggestion (McCutchan et al. submitted) that increased variance on family size might result in increased efficiencies for low h^2. They found that for 100 families of average size 10, the design with var(fs) = 7 based on a binomial distribution with mean 10 was 2% less efficient at $h^2 = 1.0$ than the equal family size design. The variable family size design became more efficient at $.25 > h > .21$ than the equal family size design, having a $var(h^2)$ 6% less than that of the equal family size design at $h^2 = .02$. Variance of family size equal to 60 is examined here. The variable family size design is less efficient at $h^2 = 1.0$, 21% higher $var(h^2)$, and more efficient at $h^2 = .02$, 35% lower $var(F^2)$, than the equal family size design.

In addition to the 100% survival case studied by McCutchan et al. (submitted), the effects of 10%, 20% and 40% random loss on this comparison are given (Table 8). The variable family size design is better buffered to loss at h ritabilities other than 1.0. The buffering is such that with 40% loss at h = .33, the variable family size design becomes more efficient.

Table 7.--Design efficiency ratios as affected by variance on the family size for 100% survival, 40 blocks, STP, 100 families and n=1000

h^2	Variance of Family Size 0 $var(\hat{h}^2)$	Variance of Family Size 60 $E_{V=60}$ 1/
1.00	.0219	1.2067
.33	.0096	1.1052
.18	.0068	.9596
.02	.0041	.6534

Table 8.--Design efficiency ratios as affected by variance on the family size for h^2 = .33, 40 blocks, STP, 100 families and n=1000

Survival (%)	Variance of Family Size 0 $var(\hat{h}^2)$	Variance of Family Size 60 $E_{V=60}$
100	.0096	1.1052
90	.0109	1.0749
80	.0128	1.0432
60	.0195	.9633

1/ $E_{V=60}$ is the ratio of the $var(\hat{h}^2)$ from the variable family size design (with variance equal to 60) divided by that from the equal family size design.

The variable family size effect on design efficiency is an extended version of the family size effect. Use of variation on the family size effectively increases the average family size through an asymmetric effect of the larger families. The deliberate use of variable family size can be viewed, consequently, in a similar light as family size, in that its use depends upon the portion of the range of h in which interest in estimation lies. If family sizes are unequal due to differential fecundity or survival, then for an average family size a variable family size design will actually be more efficient at the lower range of h^2 than an equal family size design.

A general comment can be made concerning n=1000 and n=500 designs in relationship to the criterion. Only at the 60% survival level for either h = .18 and/or h^2 = .02 are $var(h^2)$ values from the 1000-observation designs greater than the criterion. However, values from the 500-observation deigns are generally greater than the criterion for all levels of survival at h = .33, .18 and .02.

As an example of evaluating a given design for the estimation of h2, h^2 and $var(h^2)$ are estimated from a North Carolina Forest Service installed American sycamore (Platanus occidentalis L.) mother tree trial. The experiment, located in McDowell County, N.C., has 7 blocks, 30 half-sib families in 10-tree row plots and a total of 1866 surviving trees (89% survival). Variance components were estimated on eight-nar-old height data (ft.), converging in one iteration: h^2 = .25 and $\hat{var}(h^e)$ = .01. The estimated variance on h is less than that suggested by the standard, namely .016. The results of the reearch show that for a design established primarily for the estimation of an equivalent level of var(h) can be obtained with half as many total observations as in this trial.

CONCLUSIONS

Utility of efficient Modified Maximum Likelihood estimators is afforded by recent computational methodology. Both balanced and unbalanced data can be analyzed for variance component estimation in a unified approach. Design efficiencies are evaluated for the estimation of heritability using this methodology and assuming knowledge of the variance components. Rules for randomized block design allocation are established based on using the same set of variance components regardless of block, plot or family size. Single-tree plots in large blocks are recommended if the plots within blocks have small homogeneous variances--smaller blocks if the above is not possible. Recommendation of a particular family size depends on the portion of h^2 range in which estimation interest lies. Five hundred observations are insufficient to achieve the set standard on estimating h^2. One thousand observations will achieve this standard if survival is at least 80%. The rules indicate that there is not one design allocation which will uniformly provide a low var(h2) across the range of h^2.

The research has based design efficiency on the estimation of h^2. Heritability is but one function of the variance components; the methodology is laid out for the examination of other functions of variance components. This sort of a priori examination of design efficiency offers the experimenter a strong tool in achieving experimental design objectives.

LITERATURE CITED

Anderson, R. L. 1975. Designs and estimators for variance components. In: A Survey of Statistical Design and Linear Models. J. N. Srivastava (Ed.). North-Holland Pub. Co., pp. 1-29.

Anderson, R. L. 1981. Recent developments in designs and estimators for variance components. In: Statistics and Related Topics. M. Csorgo, D. A. Dawson, J. N. K. Rao, A. K. Md. E. Saleh (Eds.). North-Holland Pub. Co., pp. 3-22.

Giesbrecht, F. G. 1983. An efficient procedure for computing MINQUE of variance components and generalized least squares estimates of fixed effects. Commun. Statist.-Theor. Meth. 12(18):2169-2177.

Henderson, C. R. 1953. Estimation of variance and covariance components. Biometrics 9:226-252.

McCutchan, B. G. 1985. Design efficiencies with planned and unplanned unbalance for the estimation of heritability in forestry. Ph.D. dissertation, North Carolina State Univ., Raleigh, NC. 177 pp.

McCutchan, B. G., J. X. Ou and G. Namkoong. A comparison of planned unbalanced designs for estimating heritability in perennial crops. Submitted to Theoretical and Applied Genetics.

Namkoong, G. 1981. Variance component estimation. XVII IUFRO World Congress, Div. 6, Japan. pp. 149-159.

Namkoong, G. and J. H. Roberds. 1974. Choosing mating designs to efficiently estimate genetic variance components for trees. Silvae Genet. 23(1-3): 43-53.

Weir, R. J., Director, North Carolina State University-Industry Cooperative Tree Improvement Program, NCSU, Raleigh, NC. 1985. Personal communication.