Fixed and Random Contemporary Groups

P. M. VISSCHER and M. E. GODDARD
Department of Food and Agriculture
Livestock Improvement Unit
PO Box 500
East Melbourne
Melbourne 3002, Australia

ABSTRACT

For genetic evaluation of dairy cattle, contemporary group sizes are often small, and information is lost when contemporary groups are treated as fixed. Treating contemporary groups as random recovers some information across contemporary groups but may cause bias in prediction of breeding values if a nonrandom association exists between sires and contemporary groups. Results from a recent study indicated that treating contemporary groups as random gave consistently higher correlations between true and predicted breeding values, in particular when a positive association existed between contemporary group and sire effects. The present study shows that those results cannot be generalized based on correlations between true and predicted breeding values and on biases in breeding values for random and nonrandom association between sires and contemporary groups. For nonrandom association, the 50% best sires were assumed to be associated with half of the contemporary groups. If the best sires were used in the best contemporary groups, accuracies were higher when contemporary groups were treated as random effects (called the random model) than when contemporary groups were treated as fixed effects (fixed model) because of a reduced bias and a larger number of effective progeny. However, if the best sires were only represented in the worst contemporary groups, the correlation between true and predicted breeding values for the random model could become negative. If a nonrandom association exists between sires and contemporary groups, the groups should be treated as fixed effects for practical genetic evaluations.

(**Key words**: genetic evaluation, fixed and random models, contemporary groups, bias in predicted breeding values)

Abbreviation key: **CG** = contemporary group, **MSE** = mean squared error, **PEV** = prediction error variance.

INTRODUCTION

For genetic evaluation of dairy cattle using BLUP to predict breeding values, the main environmental effect in the linear model is often a contemporary group (**CG**) effect. In dairy cattle animal model evaluation, for example, herd-year-seasons or management groups [e.g., (9)] are usually fitted as the main environmental effect and usually treated as fixed effects. Although this treatment leads to a loss of information, in theory, predictions of the random genetic effects are unbiased (3), even if animals (sires) are nonrandomly used across CG. An example of the loss of information is that a CG in which all animals are half sibs does not contribute any information when a sire model is used. This loss of information can be reduced when CG is treated as a random effect. However, if sires are not randomly distributed over CG effects, predicted breeding values may be biased if CG is treated as a random effect (3).

For cattle breeding, CG sizes are often small. In that case, the decision to treat CG as fixed or random depends on the tradeoff between accuracy and bias. A measure to quantify accuracy of prediction and bias is the mean squared error (**MSE**): MSE = PEV + bias^2, where **PEV** is the prediction error variance. In practice, all three quantities depend on

Received August 7, 1992.
Accepted December 14, 1992.

unknown population parameters and design of data, so a general strategy to minimize MSE is not possible. Henderson (3, page 22) commented: "But in situations with small herd size I would be inclined to accept some bias in order to reduce variance." In a recent paper (7), accuracy and bias of prediction of sires' breeding values were investigated for various population parameters and for treatment of CG as either fixed or random using simulation. However, results from those studies (7) are not easy to interpret. For example, Ugarte et al. (7) found that, if a nonrandom (positive) association exists between sires and CG, a model treating CG as random was always superior for correlation between true and predicted breeding values to a model treating CG as fixed. Therefore, the introduction of bias resulted in a larger genetic gain, and, although not the conclusion of Ugarte et al. (7), it suggests that CG should be treated as random effects if evidence exists for a positive association between sires and CG. Furthermore, results were confounded by the assumption of fixed effects of phenotypic and sire variances, so a larger CG variance implied smaller residual variance and, hence, larger (within-CG) heritability. Perhaps the sire and residual variances are more logically assumed to be constant, so that within-CG heritabilities are independent of CG variances.

The objectives of this paper were 1) to show that, if a random association exists between sires and CG and if the design is reasonably balanced, simple expressions can be derived to predict accuracies of selection for models that treat CG either as fixed or random; 2) to give examples of accuracies and biases of prediction if sires are only used in certain CG.

MATERIALS AND METHODS

The notation of Ugarte et al. (7) is followed as closely as possible. Consider the model $y_{ijk} = \beta_i + u_j + e_{ijk}$, where y_{ijk} is the record on progeny k of sire j in CG i. Each of s sires has n progeny, and each of f CG has group size N. E(y) = 0, and variances and their ratios are $var(\beta) = \sigma^2_h$, $var(u) = \sigma^2_s$, $var(e) = \sigma^2_e$, $\lambda = \sigma^2_e/\sigma^2_h$, and $\alpha = \sigma^2_e/\sigma^2_s$. If CG are fixed, $\sigma^2_h = \infty$ and $\lambda = 0$. Normality of random variables is assumed throughout.

The model of genetic evaluation is $\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \mathbf{e}$, where $\mathbf{y}$, β, $\mathbf{u}$, and $\mathbf{e}$ are vectors of observations, CG effects, sire effects, and errors, respectively, and $\mathbf{X}$ and $\mathbf{Z}$ are design matrices. Solutions for sire effects may be written as

$$\hat{\mathbf{u}} = (\mathbf{Z'MZ} + \alpha\mathbf{I})^{-1}\ \mathbf{Z'My} = \mathbf{C}^{-1}\mathbf{Z'My}$$

with

$$\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X'X} + \mathbf{I}\lambda)^{-1}\mathbf{X'} = \mathbf{I} - \mathbf{XX'}/(N + \lambda),$$

because $\mathbf{X'X}$ is diagonal with element N on each diagonal. If CG are considered to be fixed in the genetic evaluation, called the fixed model, $\lambda = 0$, and for models that consider CG to be random, the random model, the variance ratio λ is added to the diagonals of the part of the mixed model equations (3) pertaining to CG effects. Diagonal elements of matrix $\mathbf{Z'MZ}$ are called effective number of progeny (daughters) and reflect the information available to predict the sire's breeding value after correction for other effects (in this case, CG effects). If $\mathbf{C}$ can be written in the form $a\mathbf{I} + b\mathbf{J}$, where $\mathbf{J}$ is a matrix of ones and a and b are scalars, then its inverse is known [e.g., (6)]:

$$\mathbf{C}^{-1} = 1/a\ [\mathbf{I} - \mathbf{J}b/(a + kb)] \qquad [1]$$

where k is the dimension of matrices $\mathbf{I}$ and $\mathbf{J}$ (and $\mathbf{C}$). Diagonal elements of $\mathbf{C}^{-1}$, d, are a function of the PEV if the model fitted is the correct model: $PEV_i = d^{ii}\sigma^2_e$, where PEV_i is the PEV, and d^{ii} is the diagonal element of $\mathbf{C}^{-1}$ for sire i. For fixed models and for random models if sires are randomly distributed over CG, PEV $= (1 - r^2)\sigma^2_s$, where r is the correlation between true and predicted breeding values or the accuracy of selection. Hence, $d = (1 - r^2)/\alpha$.

Random Association Between Sires and CG

In a completely balanced design, each sire has n/f = N/s progeny in each of the f CG. Then,

$$\mathbf{C} = \mathbf{I}(n + \alpha) - \mathbf{J}(n/s)(N/\{N + \lambda\}). \qquad [2]$$

Dimensions of $\mathbf{C}$, $\mathbf{I}$, and $\mathbf{J}$ are s. The number of effective daughters, n_e, is

$$n_e = n\{1 - (1/s)N/(N + \lambda)\}, \quad [3]$$

which reduces to $n(1 - 1/s)$ when CG are considered to be fixed. The inverse of the coefficient matrix (**C**) is, using Equation [1],

$$\mathbf{C}^{-1} = (n + \alpha)^{-1}\left[\mathbf{I} + \frac{\mathbf{J}(b_h n/s)}{n\{1 - b_h\} + \alpha)}\right]$$

$$= (n + \alpha)^{-1}\left[\mathbf{I} + \frac{\mathbf{J}(1/s)(b_h b_u)}{\{1 - b_h b_u\}}\right] \quad [4]$$

where $b_h = N/(N + \lambda)$, and $b_u = n/(n + \alpha)$. Therefore, for a balanced design, an exact prediction of the accuracy of selection and the PEV can be given using Equation [4].

If the number of sires is larger than the CG size, i.e., $N < s$ or $n < f$, the design is, by definition, unbalanced. However, an incomplete block design can be used that is partially balanced by letting N/m sires be represented in each CG with m offspring each and by letting each pair of sires occur $n(N - m)/[m^2(s - 1)]$ times in the same CG. (A design in which the progeny of each sire are randomly distributed over CG approximates this incomplete block design with $m = 1$ when $N < s$ and approximates the fully balanced design when $N > s$.) Using the incomplete block design, **C** can still be written in the form (a**I** + b**J**). Assuming that $m = 1$, diagonals of **Z'XX'Z** are n, off-diagonals are $n(N - 1)/(s - 1)$,

$$\mathbf{Z'MZ} = \mathbf{I}n\left[1 - \frac{1}{(n + \lambda} + \frac{(N - 1)}{(s - 1)(N + \lambda)}\right] - \mathbf{J}\left[\frac{n(N - 1)}{(s - 1)(N + \lambda)}\right] \quad [5]$$

and $n_e = n[1 - 1/(N + \lambda)]$.
Using Equation [5], $\mathbf{C}^{-1}$ can be obtained as before.

The simplest prediction of the accuracy of selection and PEV is to use the number of effective daughters and to predict accuracy as

$$r \approx [n_e/(n_e + \alpha)]^{.5} \quad [6]$$

and $PEV = (1 - r^2)\sigma_s^2$, using r from Equation [6]. This prediction is similar to ignoring off-diagonals of **Z'MZ**.

Nonrandom Association Between Sires and CG

If sires are nonrandomly associated with CG, the design is, by definition, unbalanced, and simple predictions of bias and accuracy of selection are not straightforward. The largest effect of unequal sire usage over CG is when some CG only contain progeny from a group of sires, and those sires have no progeny in other CG so that groups of sires and CG are completely confounded. Although this scenario is unrealistic, it clearly demonstrates the effect of treatment of CG as random. Consider a population subdivided into equally sized subsets of sires and CG. Then, if within the subset or group of CG the design is assumed to be balanced, previous results (Equations [1] to [5]) can be used because **C** will be block diagonal, and, again, its inverse is known.

Let the whole population be subdivided into p groups, so that each group contains $s^* = s/p$ sires and $f^* = f/p$ CG. Within each group, the number of effective progeny is, as before, $n_e = n(1 - N/(s^*\{N + \lambda\}))$ if $N > s^*$ and for the incomplete block design $n_e = n(1 - 1/(N + \lambda))$ if $N < s^*$. However, because the model used for analyses is no longer correct, the accuracy of selection cannot be obtained from $\mathbf{C}^{-1}$. Instead, explicit solutions for $\hat{u}$ are needed so that expectations of $v(\hat{u})$ and $cov(u,\hat{u})$ can be obtained. For a completely balanced design within groups, BLUP solutions for sire effects in group w (we [1,2, . . . p]) can be written as

$$\hat{u}_w = b_u[\bar{y}_{uw} - b_h\left[\frac{(1 - b_u)}{(1 - b_h b_u)}\right]\bar{y}_w]$$

$$= b_u[\bar{y}_{uw} - b_r\bar{y}_w] \quad [7]$$

where $b_r = [b_h(1 - b_u)]/[1 - b_h b_u]$; $\bar{y}_{uw}$ and $\bar{y}_w$ are the progeny mean for a particular sire in group w and the group mean, respectively. The group mean is equal to the mean of the progeny means and is equal to the mean of the CG within subset w. Using

$$E(\bar{y}_{uw}) = E(u_w) + E(\beta_w)$$

and

$$E(\bar{y}_w) = E(\bar{u}_w) + E(\beta_w),$$

it follows that

$$E(\hat{u}_w) = b_u[E(\bar{u}_w) + E(\beta_w)](1 - b_r), \quad [8]$$

where $E(u_w)$ means expectation for a particular group, i.e., a fixed value of w, and $E(\hat{u}_w)$, $E(\beta_w)$ are similarly defined. For fixed models, Equation [8] is zero, because $b_h = 1$ (and, therefore, $b_r = 1$). Hence, for fixed models the bias in predicted breeding values, bias = $E(u_w) - E(\hat{u}_w) = E(u_w)$. This bias is not equal to zero if $E(u_w) \neq 0$ because the information on which sires are selected (for example, true breeding values) is not included in the analysis (4). Bias = 0 if

$$b_u[E(u_w) + E(\beta_w)](1 - b_h) = E(u_w),$$

but this result is a coincidence that occurs because the parameters α, λ, n, and N "match" the sire and CG means.

For the incomplete block design, solutions for sires are

$$\hat{u}_w = nd(\bar{y}_{uw} - b_h\bar{y}_{\beta w}) + nos(1 - b_h)\bar{y}_w \quad [9]$$

with scalars d and o obtained from $\mathbf{C}^{-1} = d\mathbf{I} - o\mathbf{J}$ (see Equations [1] and [5]), and $\bar{y}_{\beta w}$ is the mean of the CG in which a particular sire is represented by its progeny. For the complete balanced case, Equation [9] reduces to Equation [7]. From Equation [9] it follows that

$$E(\hat{u}_w) = n(1 - b_h)(d + os)(E(u_w) + E(\beta_w)). \quad [10]$$

Using these results, the bias and accuracy of sire selection can be predicted for a particular design and population parameters. For each group, the expectation of u and β depend on the selection intensity applied on CG and sires, because both sires and CG may be a selected group. This prediction ignores the question how such a selection scheme may operate in practice (see Discussion). Within each group, the accuracy can be predicted if reduction in variance is taken into account using standard truncation selection theory. Calculation of the correlation between true and predicted breeding values within and across groups is shown in Appendix 1 for the case of two groups.

In the case of a subdivided population, the distribution of predicted breeding values can be regarded as a mixture of distributions about p different means. The correlation between true and predicted breeding values across the subgroups, assuming the existence of a single overall population, depends on $var(\hat{u}_w)$ and $cov(u_w,\hat{u}_w)$ within all groups and on the variance of group means across all groups (Appendix 1).

The simplest example is when the best 50% of sires are used in the best 50% of CG. Hence, two groups of CG exist, and s/2 sires have all n progeny in f/2 CG. Selection intensities (i) for both CG and sires are .8, and the reduction in variance (1 - k) for each of the two groups is .64 so that $E(u_w)$ and $E(\beta_w)$ are $i\sigma_s$ and $i\sigma_h$ in the group containing the top sires and CG and are $-i\sigma_s$ and $-i\sigma_h$ in the other group. When a fixed model is used for genetic evaluation, the expectation of predicted breeding values in each group is zero, so the bias within each group is $\pm i\sigma_s$. A consequence of using a random model is that the estimate of the difference in mean predicted breeding values between the top and bottom group is changed. For this example, the variance of predicted breeding values across the conceptual total population is simply the sum of the variance within groups and the square of the expectation of the predicted breeding values within groups (see Appendix 1).

Random Model with Overall Mean

The random models in the previous sections implicitly assumed that the mean was known, because no overall mean was fitted in the model of evaluation. Hence, following Ugarte et al. (7), both location and dispersion parameters were assumed to be known for the random model. The question arises whether results would be different if an overall mean were fitted in the random model, because for practical genetic evaluation at least one fixed effect would be fitted. Appendix 2 shows that the effect of fitting an overall mean is relatively small.

RESULTS

Random Association Between Sires and CG

When the same range of values was used for N and λ as Ugarte et al. (7), accuracies of selection and PEV were predicted using Equa-

TABLE 1. Predicted accuracy of selection (r^1) for different contemporary group (CG) sizes (N) with CG analyzed as fixed effect (F) or random effect (R) when sires were randomly used across CG and sire and phenotypic variance were constant.

	Ratio of residual to CG variance							
	1.5		2.75		5.25		17.75	
N	F	R	F	R	F	R	F	R
2	.826	.872	.799	.861	.779	.856	.761	.853
3	.859	.880	.836	.866	.819	.858	.803	.853
4	.871	.884	.850	.869	.834	.859	.819	.854
6	.881	.889	.862	.872	.847	.862	.832	.854
9	.887	.892	.868	.875	.854	.863	.840	.855
12	.890	.894	.871	.877	.857	.864	.844	.855
18	.893	.895	.874	.878	.861	.866	.847	.855
24	.894	.896	.876	.879	.862	.866	.849	.856
48	.896	.897	.878	.879	.864	.867	.851	.856
120	.896	.896	.878	.879	.865	.866	.852	.854
240	.896	.896	.878	.879	.865	.865	.852	.853
$h_w^{2\ 2}$	.400	.400	.333	.333	.294	.294	.263	.263

[1]$r = [1 - PEV/\sigma_s^2]^{.5}$, using Equations [1], [4], and [5].

[2]Within-CG heritability = $.25/(1 - \sigma_h^2/\sigma_y^2)$.

tions [4] and [5]. In cases for which $n(N - 1)/(s - 1)$ is not an integer and $N < s$, the exact incomplete block design is not possible, and the results are an approximation. In all subsequent predictions, $n = 40$, and $s = 60$; hence, $Nf = ns = 2400$. If the phenotypic and sire variances are held constant, as in Ugarte et al. (7), then $\alpha = \alpha_0 \lambda/(\lambda + 1)$, where α_0 is the ratio of residual to sire variance if the CG variance is zero. The within-CG heritability ($h_w^2 = 4/(1 + \alpha)$) is equal to $h_0^2/(1 - c^2)$, where h_0^2 is the heritability ($= 4\sigma_s^2/\sigma_y^2$), which is .25 in all cases, and c^2 is the CG variance as a proportion of the phenotypic variance. In Table 1, predicted accuracies assume a constant phenotypic variance. Table 1 can be directly compared with Table 3 of Ugarte et al. (7). Results, as expected, are nearly identical, showing that the balanced designs approximate well the random distribution of progeny across CG. The last row in Table 1 shows the within-CG heritability for a particular value of λ. A more relevant comparison is to allow the phenotypic variance to increase so that α is constant. For this case, results for the same range of parameters are shown in Table 2; now only one column is needed for the fixed effects model.

The prediction of accuracies and PEV using Equation [6] using number of effective progeny gave similar results (not shown) with accuracies approximately .01 too high. The number of effective daughters is also shown in Table 2. Table 2 clearly shows that the advantage of using prior information about the distribution (variance) of CG is largest for small CG sizes. For $N > 24$, accuracies scarcely changed for the range of parameters used. Even for $N = 4$, the gain in accuracy was .031 or less.

Nonrandom Association Between Sires and CG

For the example of two subpopulations each with 50% of sires and CG, three different scenarios were considered. 1) Both sires and CG were selected at random. 2) Sires and CG were selected using truncation selection; the best sires were used in the top CG. 3) Sires and CG were selected using truncation selection, but top sires are only used in the worst CG. Scenario 2 corresponds to the design of Ugarte et al. (7). Scenarios 2 and 3 represent a special form of selection because selection is based on true breeding values and true CG effects [see Henderson (4) for a detailed description and discussion of this kind of

TABLE 2. Predicted accuracy of selection (r^1) and number of effective progeny (n_e^2) for different contemporary group (CG) sizes (N) with CG analyzed as fixed effect (F) or random effect (R) when sires were randomly used across CG and residual and sire variance were constant.

			Ratio of residual to CG variance for R							
	F		1.50		2.75		5.25		17.75	
N	r	n_e	r	n_e	r	n_e	r	n_e	r	n_e
2	.752	20.0	.809	28.6	.823	31.6	.835	34.5	.847	38.0
3	.796	26.7	.821	31.1	.829	33.0	.837	35.2	.847	38.1
4	.812	30.0	.827	32.7	.833	34.1	.839	35.7	.847	38.2
6	.826	33.3	.834	34.7	.837	35.4	.841	36.4	.848	38.3
9	.834	35.6	.838	36.2	.840	36.6	.843	37.2	.848	38.5
12	.837	36.7	.841	37.0	.842	37.3	.844	37.7	.849	38.9
18	.841	37.8	.843	37.9	.844	38.1	.846	38.3	.849	38.9
24	.843	38.3	.844	38.4	.845	38.5	.846	38.6	.850	39.0
48	.845	39.2	.846	39.2	.846	39.2	.847	39.2	.851	39.4
120	.846	39.3	.846	39.3	.846	39.3	.847	39.4	.851	39.4
240	.846	39.3	.846	39.3	.846	39.3	.846	39.3	.851	39.4

[1]$r = [1 - PEV/\sigma_s^2]^{.5}$, using Equations [1], [4], and [5].

[2]Diagonal element of Z'MZ, using Equations [3] and [5].

selection]. Sire and residual variances were 625 and 9375, respectively. Predicted accuracies within and across groups, biases, and MSE are shown in Tables 3 and 4 for these three scenarios. Results in Table 3 are for relatively large CG sizes (N ≥ 30), using Equations [7] and [8], and in Table 4 for small CG sizes (N ≤ 12), using Equations [9] and [10]. For the case considered herein, the average bias in predicted breeding value across the two subpopulations is zero for both fixed and random models, but the bias within each subpopulation is more informative; this bias is given in Tables 3 and 4.

TABLE 3. Predicted bias,[1] mean squared error (MSE), correlation between true and predicted breeding values within and across groups (r_w and r_b)[2] for fixed (F) and random (R) models for large contemporary groups (CG) sizes when the population is subdivided into two separate groups each containing 50% of sires CG.[3,4]

		F				R (λ = 1.50)				R (λ = 17.75)			
Scenario[5]	N	r_w	r_b	bias	MSE	r_w	r_b	bias	MSE	r_w	r_b	bias	MSE
1	30	.838	.838	0	186	.841	.841	0	183	.848	.848	0	175
	60	.838	.838	0	186	.840	.841	0	184	.846	.846	0	178
	120	.838	.838	0	186	.839	.839	0	185	.844	.844	0	180
	240	.838	.838	0	186	.839	.839	0	185	.842	.842	0	182
2	30	.690	.415	20	542	.692	.771	11	255	.698	.885	1	142
	60	.690	.415	20	542	.691	.645	15	365	.697	.852	5	172
	120	.690	.415	20	542	.690	.545	17	443	.695	.781	10	245
	240	.690	.415	20	542	.690	.484	19	490	.693	.677	14	339
3	30	.690	.415	20	542	.692	.157	25	759	.698	.461	19	508
	60	.690	.415	20	542	.691	.275	23	653	.697	.451	19	516
	120	.690	.415	20	542	.690	.343	21	598	.695	.440	20	524
	240	.690	.415	20	542	.690	.379	21	570	.693	.430	20	531

[1]$E(u_w) - E(\hat{u}_w)$ pertaining to the group with the best sires.

[2]Appendix 1 shows calculations.

[3]$\sigma_s^2 = 625$, $\sigma_e^2 = 9375$.

[4]λ is ratio of residual to CG variance.

[5]Scenario 1: Sires and CG randomly distributed across the two groups; scenario 2: best sires and best CG in one group; scenario 3: best sires and worst CG in one group.

TABLE 4. Predicted bias,[1] mean squared error (MSE), correlation between true and predicted breeding values within and across groups (r_w and r_b)[2] for fixed (F) and random (R) models for small contemporary group (CG) sizes when the population is subdivided into two separate groups each containing 50% of sires CG.[3,4]

		F				R (λ = 1.50)				R (λ = 17.75)			
Scenario[5]	N	r_w	r_b	bias	MSE	r_w	r_b	bias	MSE	r_w	r_b	bias	MSE
1	2	.749	.749	0	275	.809	.809	0	216	.847	.847	0	177
	4	.807	.807	0	218	.826	.826	0	199	.847	.847	0	176
	9	.829	.829	0	196	.836	.836	0	189	.848	.848	0	176
	12	.832	.832	0	192	.838	.838	0	187	.848	.848	0	175
2	2	.568	.342	20	592	.661	.883	−24	745	.698	.901	−7	191
	4	.644	.388	20	562	.672	.893	−15	377	.697	.901	−6	181
	9	.675	.407	20	548	.684	.892	−3	156	.698	.899	−5	162
	12	.680	.410	20	546	.685	.880	1	147	.698	.898	−4	154
3	2	.568	.342	20	592	.661	−.453	43	1998	.698	.478	19	495
	4	.644	.388	20	562	.672	−.350	38	1604	.697	.476	19	497
	9	.675	.407	20	548	.684	−.162	32	1161	.698	.472	19	500
	12	.680	.410	20	546	.686	−.082	30	1034	.698	.470	19	501

[1]$E(u_w) - E(\hat{u}_w)$ pertaining to the group with the best sires.

[2]Appendix 1 shows calculations.

[3]$\sigma_s^2 = 625$, $\sigma_e^2 = 9375$.

[4]λ is ratio of residual to CG variance.

[5]Scenario 1: Sires and CG randomly distributed across the two groups; scenario 2: best sires and best CG in one group; scenario 3: best sires and worst CG in one group.

Because CG sizes in Table 3 are ≥ 30, accuracies within groups are almost the same for fixed and random models. Therefore, the differences in accuracies between groups and MSE are almost entirely due to differences in bias. For scenario 2, the random models give lower MSE and bias and higher accuracies than the fixed models. Although results from Table 3 cannot be compared directly with results from Ugarte et al. (7), these trends for scenario 2 (positive association between sires and CG) are similar to their results. Table 3 shows that, for particular sets of parameters, biases in random effects models can be zero. For example, for λ = 17.75 and N = 30, the bias was approximately zero (Table 3), as expected from Equation [8].

Differences in accuracies and MSE between fixed and random models when CG sizes are small (Table 4) are a result of different amounts of information (n_e larger for random models) and differences in bias. For scenario 2, very high accuracies were obtained, similar to those reported by Ugarte et al. (7). However, for scenario 3, i.e., a negative association between sires and CG, correlation between true and predicted breeding values was sometimes negative. Hence, for very small CG sizes and random models, the average predicted breeding values of the top sires can be smaller than the average predicted breeding values of the worst sires. As shown in Table 4 for scenario, biases can be negative; i.e., $E(\hat{u}_w) > E(u_w)$.

For some combinations of parameters, the correlation between true and predicted breeding values across groups were larger than the maximum accuracy in the absence of an association between sires and CG, as also reported by Ugarte (7) (see Table 2 and values pertaining to scenario 1 in Tables 3 and 4). Their explanation of this phenomenon was that the variance of predicted breeding values was increased. The reason for this is the contribution of the between CG variance to the variance of predicted breeding values (see Equations [7] to [10]). Hence, a better explanation of this phenomenon might be that the environmental differences between CG are used to predict genetic differences between sires.

In scenario 3, the bias for the random model was sometimes greater than for the fixed effect model, particularly when CG sizes were small

(Table 4). In these cases, the accuracies are lower, and MSE are higher, for random than for fixed models. The largest bias, for λ = 1.50, results in selection of sires from the subpopulation with the worst sires, so the correlation between true and predicted breeding values becomes negative (Table 4).

DISCUSSION

If the assumption of a random association between sires and CG is valid in practice, then CG should be treated as random effects for the practical genetic evaluation because more information is used to predict breeding values, so PEV decreases, and accuracy of selection increases. As quantified in this study for a range of population parameters, the effective number of progeny increases when CG are random, or, more formally, when interblock information is recovered. The gain is marginal, however, unless CG sizes are $\leq$ 4. Alternative ways of recovering information between CG of small sizes were proposed by Chauhan and Thompson (2), Chauhan (1), Wade et al. (8), and Schmitz et al. (5).

In the case of a nonrandom association between sires and CG, the accuracy of selection can be either increased or decreased by the use of random CG. In our calculations and in those of Ugarte et al. (7), a random model increases accuracy largely by reducing the bias for the fixed CG model. This result is the opposite of the conventional idea that random CG models decrease PEV but cause bias. In practice, the bias for fixed CG models is eliminated in other ways. Nonrandom association of sires and CG is not a biological property of the data but represents a form of selection (4). The nature of this selection must be considered. 1) If selection is based on predicted breeding values and if the corresponding data are included in the genetic evaluation, then predicted breeding values are unbiased if CG are fixed. This property is a standard result of BLUP theory (3). Treatment of CG as random will produce biased breeding values. 2) If the nonrandom association is the result of outside information, for example, information about mean genetic differences between countries, this information amounts to knowledge about groups. Ignoring (fixed) groups in the model then underestimates group differences, particularly if linkage is poor between groups. For the extreme example given herein, in which two unlinked groups each represented half of the sires and CG, the estimate of the group difference was zero, whereas the real difference was $2i\sigma_s$. A model with fixed CG fitting sires within the two groups would give unbiased predicted breeding values.

Treatment of CG as random in the presence of nonrandom association between CG and sires uses nonrandom environmental differences between CG to estimate nonrandom differences between sires. If this treatment improves the prediction of breeding values, as reported by Ugarte et al. (7) and shown in the example in this study, it is fortuitous. In practical sire evaluation schemes, to attempt to use nonrandom CG differences to estimate sire breeding values would be quite unacceptable. For example, what would happen if the best bulls were used in the worst CG? Treatment of CG as random would then decrease the accuracy of sire evaluation, as shown for scenario 3. This result was also discussed by Ugarte et al. (7). Another example of the potential negative effect of treating CG as random is when selected (proven) bulls are used in CG with different means than CG (herds) in which young bulls are progeny tested, because then the difference in predicted breeding values between the young and old bulls would not reflect the true difference in breeding values between them. A final example is the case in which this year's progeny-test team of bulls have daughters in relatively low CG because of adverse environmental effects (e.g., a drought) because then the comparison of bulls across years would be biased. In general, the use of the term "correlation" between sires and CG can be misleading because this correlation has no biological interpretation and is not a population parameter in the usual sense: a nonrandom association between sires and herds in practice is merely the result of farmers choosing sires, and this association should not be used to predict the performance of future progeny of the sires.

CONCLUSIONS

When there is clearly no association between sires and CG, CG may be treated as random; otherwise, CG should be treated as

fixed. If nonrandom association between sires and CG exists because of a group effect, this effect should be fitted in the model.

ACKNOWLEDGMENTS

This research was supported by a grant from the Australian Dairy Research and Development Corporation.

REFERENCES

1 Chauhan, V.P.S. 1987. Dairy sire evaluation fitting some of the herd-year-season effects as random. Livest. Prod. Sci. 16:117.

2 Chauhan, V.P.S., and R. Thompson. 1986. Dairy sire evaluation using a "rolling months" model. J. Anim. Breed. Genet. 103:321.

3 Henderson, C. R. 1973. Sire evaluation and genetic trends. Page 10 *in* Proc. Anim. Breeding Genet. Symp. in Honor of Dr. J. L. Lush, Am. Soc. Anim. Sci., Am. Dairy Sci. Assoc., Champaign, IL.

4 Henderson, C. R. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423.

5 Schmitz, F., R. W. Everett, and R. L. Quaas. 1991. Herd-year-season clustering. J. Dairy Sci. 74:629.

6 Searle, S. R. 1982. Matrix Algebra Useful for Statistics. John Wiley & Sons, New York, NY.

7 Ugarte, E., R. Alenda, and M. J. Carabano. 1992. Fixed or random groups in genetic evaluations. J. Dairy Sci. 75:269.

8 Wade, K. M., R. L. Quaas, and L. D. Van Vleck. 1990. Mixed linear models with an autoregressive error structure. Proc. 4th World Congr. Genet. Appl. Livest. Prod., Edinburgh, Scotland XIII:508.

9 Wiggans, G. R., I. Misztal, and L. D. Van Vleck. 1988. Implementation of an animal model for genetic evaluation of dairy cattle in the United States. J. Dairy Sci. 71(Suppl. 2):54.

APPENDIX 1

Correlation Between True and Predicted Breeding Values in Subdivided Population

The population is subdivided into two groups of sires and CG. Each group has s/2 sires with n progeny each, and f/2 CG with N observations each. Subscripts 1, 2, and w are used to indicate group 1, group 2, and group 1 or 2, respectively. Parameters without a subscript refer to the conceptual entire population. Sires and CG are assumed to be selected with the same standardized selection intensity (i), and, without loss of generality, group 1 is assumed to contain the best sires. Within each group, no association between sires and CG exists. Then

$$E(u_1) = -E(u_2) = i\sigma_s,\ v(u_1) = v(u_2) = v(u_w) = (1 - k)\sigma_s^2, \text{ and } v(u) = \sigma_s^2.$$

The variance of predicted breeding values for the single population is

$$v(\hat{u}) = E(\hat{u}^2) - [E(\hat{u})]^2 = E(\hat{u}^2), \quad [A1]$$

because $E(\hat{u}) = [E(\hat{u}_1) + E(\hat{u}_2)]/2 = 0$. Therefore,

$$\begin{aligned} v(\hat{u}) &= [E(\hat{u}_1^2) + E(\hat{u}_2^2)]/2 \\ &= E(\hat{u}_w^2) \\ &= v(\hat{u}_w) + [E(\hat{u}_w)]^2. \end{aligned} \quad [A2]$$

Similarly,

$$\begin{aligned} \text{cov}(u,\hat{u}) &= E(u,\hat{u}) - E(u)E(\hat{u}) \\ &= E(u_w,\hat{u}_w) \\ &= \text{cov}(u_w,\hat{u}_w) + E(u_w)E(\hat{u}_w). \end{aligned} \quad [A3]$$

The correlation between true and predicted breeding values within groups and across groups are defined as

$$r_w = [\text{cov}(u_w,\hat{u}_w)]/[v(u_w)v(\hat{u}_w)]^{.5} \quad [A4]$$

and

$$r_b = [\text{cov}(u,\hat{u})]/[v(u)v(\hat{u})]^{.5} \quad [A5]$$

To calculate r_w and r_b, expressions for $E(\hat{u}_w)$, $v(\hat{u}_w)$, and $\text{cov}(u_w,\hat{u}_w)$ are needed. Depending on the design (balanced or balanced incomplete block design), $E(\hat{u}_w)$ is calculated using Equation [8] or [10] from the main text, and $v(\hat{u}_w)$ and $\text{cov}(u_w,\hat{u}_w)$ from Equation [7] or [9].

Consider, for example, the balanced design. From Equation [7] it follows that

$$v(\hat{u}_w) = b_u^2[v(\bar{y}_{uw}) - b_r(2 - b_r)v(\bar{y}_w)],$$

which is calculated using

$$v(\bar{y}_{uw}) = (1 - k)\sigma_s^2 + (1 - k)\sigma_h^2/(f/2) + v(e)/n,$$

and

$$v(\bar{y}_w) = (1 - k)\sigma_s^2/(s/2) + (1 - k)\sigma_h^2/(f/2) + v(e)/(ns/2).$$

The covariance between true and predicted breeding values also follows from Equation [7],

$$\begin{aligned}\text{cov}(u_w,\hat{u}_w) &= b_u[\text{cov}(u_w,\bar{y}_{uw}) - b_r\ \text{cov}(u_w,\bar{y}_w)] \\ &= b_u[1 - b_r/(s/2)](1 - k)\sigma_s^2.\end{aligned}$$

Although $\text{cov}(u_w,\hat{u}_w)$, and, therefore, r_w are always positive, $\text{cov}(u,\hat{u})$ and r_b can be negative if $E(u_w)$ and $E(\hat{u}_w)$ are large and opposite of sign.

APPENDIX 2

Results for Random Models When an Overall Mean Is Fitted

Random Association Between Sires and CG. The model is now $y_{ijk} = \mu + \beta_i + u_j + e_{ijk}$; μ is the overall mean. The part of the coefficient matrix in the mixed model equations relating to μ and β is

$$\mathbf{X'X} + \lambda\mathbf{I}^* = \begin{bmatrix} Nf & N\mathbf{1}' \\ N\mathbf{1} & \mathbf{X'X} + \lambda\mathbf{I} \end{bmatrix}$$

with $\mathbf{X} = [\mathbf{1}\ \mathbf{X}]$, $\mathbf{I}^* = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I} \end{bmatrix}$, and $\mathbf{1}$ a vector of ones.
Then,

$$[\mathbf{X'X} + \lambda\mathbf{I}^*]^{-1} = (1/f\lambda)\begin{bmatrix} b_h & -\mathbf{1}' \\ -\mathbf{1} & f(1 - b_h)\mathbf{I} + b_h\mathbf{J} \end{bmatrix} \quad [B1]$$

For the completely balanced case, it is easily shown that

$$\mathbf{Z'X}(\mathbf{X'X} + \lambda\mathbf{I}^*)^{-1}\mathbf{X'Z} = (n/s)\mathbf{J}, \text{ and}$$
$$\mathbf{C} = (n + \alpha)\mathbf{I} - (n/s)\mathbf{J} \quad [B2]$$

Therefore, $n_e = n(1 - 1/s)$, which is the same as the number of effective progeny for the fixed case, and, hence, for the balanced case random CG has no advantage if an overall mean is fitted in the model. For the balanced incomplete block design, it can be shown that

$$\mathbf{C} = \mathbf{I}\,[\alpha + n(1 - s(1 - N/s)/\{(s - 1)(N + \lambda)\}] - \mathbf{J}\,[n(1/s - (1 - N/s)/\{s - 1)(N + \lambda)\}] \quad [B3]$$

From Equation [B3], it follows that $n_e = n\{1 - 1/(N\lambda) - (1 - b_h)/s\}$. This quantity is similar to the n_e that followed from Equation [5] in the main text with an additional term, $(1 - b_h)/s$, which can be interpreted as the information lost by estimation of the overall mean. Because the extra terms is of the order 1/s, the effect of including an overall mean in the random model is small provided that s is large.

Subdivided Population. We consider only the balanced case; i.e., group 1 and 2 each contain s/2 sires with n progeny distributed over f/2 CG, and each sire has 2n/f progeny in each of f/2 CG. After some tedious algebra, it can be shown that

$$\mathbf{C} = \begin{bmatrix} \mathbf{I}(n + \alpha) - x\mathbf{J} & -y\mathbf{J} \\ -y\mathbf{J} & \mathbf{I}(n + \alpha) - x\mathbf{J} \end{bmatrix} \quad [B4]$$

with $x = (n/s)(1 + b_h)$, and $y = (n/s)(1 - b_h)$.

$$\mathbf{C}^{-1} = \begin{bmatrix} a\mathbf{I} + b\mathbf{J} & c\mathbf{J} \\ c\mathbf{J} & a\mathbf{I} + b\mathbf{J} \end{bmatrix},$$

with

$$a = 1/(n + \alpha),$$

$$b = \left[\frac{(1/sn)b_u\,[(1 - b_u^2 b_h)}{(1 - b_h b_u)\,(1 - b_u) - 1}\right], \text{ and}$$

$$c = \left[\frac{(1/sn)b_u\,[b_u(1 - b_h)}{(1 - b_h b_u)\,(1 - b_u)}\right].$$

Elements of $\mathbf{Z'My}$ are $n[\bar{y}_{uw} - \bar{Y} - b_h(\bar{y}_w - \bar{Y})]$; $\bar{y}_{uw}$,$\bar{y}_w$, and $\bar{Y}$ are the progeny mean of a sire in group w (1 or 2), the group mean and the overall mean, respectively. It follows that

$$\hat{u}_w = b_u\,[\bar{y}_{uw} - \bar{Y} - b_h(\bar{y}_w - \bar{Y})] + ns^*b(1 - b_h)(\bar{y}_w - \bar{Y})$$

$$+ ns^*c(1 - b_h)(\bar{y}_{w'} - \bar{Y}) \text{ for } w \neq w'. \quad [B5]$$

Using $E(u_w) = -E(u_{w'})$ and $E(\beta_w) = -E(\beta_{w'})$, the expectation of the predicted breeding value i group w (w') reduces to

$$E(\hat{u}_w) = b_u(1 - b_r)\, E(u_w + \beta_w) \quad [B6]$$

with, as before, $b_r = [b_h(1 - b_u)]/[1 - b_h b_u]$. Equation [B6] is identical to Equation [8] of the main text. Hence, for the balanced case, the expectation of $\hat{u}_w$ and, therefore, the expectation of the bias are the same whether an overall mean is fitted or not.

We conclude that fitting an overall mean in the random model has little effect on the results and, therefore, that overall conclusions from the main text are not restricted to the case of known location parameters.