# Power of Likelihood Ratio Tests for Heterogeneity of Intraclass Correlation and Variance in Balanced Half-Sib Designs

PETER M. VISSCHER[1]
Institute of Cell, Animal and Population Biology
University of Edinburgh
West Mains Road, Edinburgh EH9 3JT, Scotland

## ABSTRACT

Statistical power of likelihood ratio tests was investigated for detection of heterogeneous variances and intraclass correlation in balanced half-sib designs. Powers of likelihood ratio tests were obtained from simulations. For half-sib designs of sires nested within herds, true intraclass correlations and phenotypic variances, and estimates thereof, were repeatedly sampled, and likelihood ratio tests were conducted. The power for detecting heterogeneity of intraclass correlations was low, but the power for detecting heterogeneous phenotypic variances was nearly always 100%. For balanced cross-classified designs, sires had progeny in all herds, and data were simulated by assuming that heterogeneity of between- and within-sire components was the result of a herd-dependent scale effect. Using this model, the power to detect heterogeneous between-sire components was substantially higher than the corresponding power to detect heterogeneous intraclass correlations in the nested design. It seems unlikely that reliable inference about heterogeneity of genetic variances or heritabilities between individual herds from daily cattle field data can be made.

(**Key words**: statistical power, likelihood ratio test, heterogeneity of variance)

**Abbreviation key:** HYS = herd-year-season, IAM = individual animal model, ICC = intraclass correlation, LR = likelihood ratio, MCPB = mean cross product for sires between strata, MLE = maximum likelihood estimate, MSB = mean square between sires within a stratum, MSW = mean square within sires within a stratum.

## INTRODUCTION

In animal breeding, BLUP [e.g., Henderson (8)] has become the method of choice for predicting breeding values from mixed linear models. Theoretically, (co)variances of random effects included in the mixed model should be known without error, but, in practice, estimates are used. It has become standard practice to estimate variances using REML [Patterson and Thompson (15)]. The most desirable (linear) model both for prediction of breeding values and estimation of genetic parameters appears to be an individual animal model (IAM), in which relationships between all animals in the data and pedigree are taken into account [e.g., (19, 25) for applications in dairy cattle].

One assumption usually made by users of BLUP is homogeneity of variances across levels of fixed (and random) effects. In dairy cattle, however, there is abundant evidence from recent analyses that this assumption is not valid (2, 3, 4, 13, 14, and 18). Typically for studies investigating heterogeneity of variance, herds or herd-year-seasons (HYS) are grouped according to their mean production or phenotypic variance, and parameters are estimated within and between groups using a sire model. Unfortunately, using an IAM for estimating parameters is computationally demanding, and relatively small sample sizes are necessarily used to estimate population parameters. One suggestion for dairy cattle parameter estimation is to use individual herd data as samples (20, 22, 23, 24) and to combine several individual herd estimates into a population estimate. Using individual herd data separately provides a framework to investigate heterogeneity of variance between herds (23, 24). If results about variance heterogeneity

Received March 18, 1991.
Accepted December 9, 1991.

[1] Present address: Department of Agriculture and Rural Affairs, Livestock Improvement Unit, 166–176 Wellington Parade, East Melbourne, Victoria 3002, Australia.

from a sample of individual herd estimates can be extrapolated to the total population (of herds), then, for any trait and parameterization in heritability and phenotypic variance, one of the following conclusions can be drawn from one such sample: 1) both heritabilities and phenotypic variances are homogeneous across herds; 2) heritabilities are homogeneous, and phenotypic variances are heterogeneous across herds; 3) heritabilities are heterogeneous and phenotypic variances are homogeneous across herds; and 4) both heritabilities and phenotypic variances are heterogeneous across herds.

The (arbitrary) parameterization in heritabilities and phenotypic variances, instead of parameterization in additive genetic and environmental variances, was chosen to investigate previously reported conclusions about heterogeneity of variance between herds (24), which were in terms of the same parameterization. Furthermore, results from estimating variances in dairy cattle are most commonly reported in terms of heritability and phenotypic variances. The implications of these four scenarios for a (national) BLUP evaluation, if the appropriate covariance structure of the data is used, vary substantially. Scenarios 2 to 4 imply that estimates for individual herds should be obtained regularly, which is tedious and may be subject to sampling error. Furthermore, in addition to (sampling) problems associated with estimation of the relevant parameters, there may be computational problems with a large-scale implementation.

Inference about the (co)variance structure of observations across herds or HYS, therefore, has implications for the choice of the desirable model to be used. What significance test should be used in selecting the most likely scenario, and how powerful are such tests for small sample sizes? Because the estimation procedure usually is REML, it seems natural to use a likelihood ratio (LR) test, which has desirable asymptotic properties (1).

The aim of this study was to investigate the power of an LR test in detecting heterogeneous variances for individual groups (herds).

An LR test was used to test whether heritability differed between herds while allowing for heterogeneous individual herd phenotypic variances. To predict the power of an LR test for a given design, the distribution of the test statistic was required. Unfortunately, the distribution of the LR from IAM-REML variance component estimation is not known. One suggestion was to investigate the detection of differences in between- and within-sire variances in different herds using balanced half-sib designs, because the distribution of the test statistic from such a design is known when ANOVA is used to estimate variances. Therefore, ANOVA theory was used in the prediction of the power. Both nested and cross-classified half-sib designs were used to contrast the statistical power in detecting heterogeneous variances across individual herds for both designs.

## MATERIALS AND METHODS

### Balanced Nested Half-Sib Designs

Suppose that there are observations on progeny of sires in different herds (or strata) and that an LR test is used to determine whether a particular set of herds differs in intraclass correlation (ICC), phenotypic variance, or both. The ICC is the ratio of between-sire variance to the sum of between- and within-sire variance and is usually assumed to be one-quarter of the heritability. It was assumed that sires were nested within herds. Let there be sn observations in each herd from s sires with n progeny each. Then, assuming normality, the log-likelihood of error contrasts (15) for data from herd i [see, for example, (21)] is, apart from a constant,

$$L_i = -\frac{1}{2}\{s(n-1)\log(\sigma^2_{wi}) + (s-1)\log(\sigma^2_{wi} + n\sigma^2_{bi}) + W_i/(\sigma^2_{wi}) + B_i/(\sigma^2_{wi} + n\sigma^2_{bi})\}$$

with

$\sigma^2_{wi}$ = within-sire variance in herd i,

$\sigma^2_{bi}$ = between-sire variance in herd i,

$W_i$ = within-sire sum of squares for herd i, and

$B_i$ = between-sire sum of squares for herd i.

Reparameterization in $t_i$, the ICC for herd i, and $\sigma^2_i$, the phenotypic variance in herd i, gives

$$\begin{aligned}L_i = -\frac{1}{2}\{ & s(n-1)\log(1-t_i) \\ & + (s-1)\log(1+(n-1)t_i) \\ & + W_i/(\sigma_i^2(1-t_i)) + B_i/(\sigma_i^2(1+(n-1)t_i)) \\ & + (sn-1)\log(\sigma_i^2)\}.\end{aligned}$$

For data from k herds, assuming that sires are nested within herds, the log-likelihood is

$$L_u = \Sigma L_i \qquad [1]$$

and the (residual) maximum likelihood is obtained by substituting the ANOVA estimates for $t_i$ and $\sigma_i^2$ in [1], for $t_i > 0$. Now consider the null hypothesis that the ICC are the same in all herds while allowing for heterogeneous phenotypic variances across herds, and let the common value of the ICC be $t_0$. Then,

$$\begin{aligned}L_0(x_{ij}|t_0,\sigma_1\ldots\sigma_k) = & \\ -\frac{1}{2}\sum_i^k & [s(n-1)\log(1-t_0) \\ & + (s-1)\log(1+(n-1)t_0) \\ & + W_i/(\sigma_i^2(1-t_0)) \\ & + B_i/(\sigma_i^2(1+(n-1)t_0)) \\ & + (sn-1)\log(\sigma_i^2)]. \qquad [2]\end{aligned}$$

The REML estimates for $t_0$ and $\sigma_i^2$ satisfy, respectively,

$$\sum_i^k \left[\frac{-s(n-1)}{(1-t_0)} + \frac{(s-1)(n-1)}{(1+(n-1)t_0)} + \frac{W_i}{\sigma_i^2\,(1-t_0)^2} - \frac{(n-1)B_i}{\sigma_i^2(1+(n-1)t_0)^2}\right] = 0 \qquad [3]$$

and

$$\frac{W_i}{(1-t_0)} + \frac{B_i}{(1+(n-1)t_0)} - (ns-1)\sigma_i^2 = 0. \qquad [4]$$

There is no explicit solution for $t_0$ and $\sigma_i^2$, and iterative techniques must be used to solve [3] and [4] and to obtain the maximum likelihood estimate (MLE). Similar formulas can be derived for the hypothesis that the phenotypic variances are homogeneous while allowing the ICC to differ between herds or for the hypothesis that both the ICC and phenotypic variances are homogeneous (standard ANOVA formulas).

The power of an LR test to detect heterogeneity of ICC or phenotypic variances was investigated by simulation. For each replicate, the true $t_i$ were sampled from a truncated normal distribution with mean $t_0$ (hence, $t_i$ was approximately $\sim N(t_0,v(t_i))$) in the interval <0,1>. Repeatedly sampling of ICC and variances can be justified because the interest is in inference about the whole population of herds, and the power of detecting different variances for a particular (arbitrary) set of herds would be conditional on the parameters for those herds. In essence, the calculated power is the expectation of the power over the distribution of ICC. For each of k herds, between- and within-sire sum of squares were sampled from the appropriate chi-square distribution, and the sample between- and within-sire components were estimated using REML. The sampling procedure caused a slightly skewed distribution of $t_i$ because $t_0$ was .1. By sampling sum of squares, data were assumed to be corrected for all fixed effects, including fixed herd effects.

Data were simulated for two different designs: the first design with k = 25, s = 30, and n = 10, and a second design with k = 10, s = 100, n = 10. For each of 5000 replicates, LR tests were carried out corresponding to the following null hypotheses ($H_0$): 1) $H_0$ $[\sigma_0^2, t_0]$ = both ICC and phenotypic variances are homogeneous (df = 2(k – 1)); 2) $H_0$ $[\sigma_i^2, t_0]$ = ICC are homogeneous, allowing for heterogeneous phenotypic variances (df = k – 1); and 3) $H_0$ $[\sigma_0^2, t_i]$ = phenotypic variances are homogeneous, allowing for heterogeneous ICC (df = k – 1). The alternative hypothesis in all cases was heterogeneity of both $\sigma_i^2$ and $t_i$. For each replicate, the appropriate REML estimates were calculated for each hypothesis using simple iterative techniques. Two different sources are expected to cause biases in the LR test: one source is that small samples cause departures of the distribution of the test statistic from the chi-square distribution; the other source is that the estimates of the ICC are not normally distributed.

To predict the power of the LR test, the sampling variances of the parameter estimates are needed, but, using REML, the exact sampling variances of the estimates are not known. One suggestion is to use approximate sampling variances pertaining to ANOVA estimates. Assuming that $\hat{t}_i$ is estimated from an ANOVA, its sampling variance (5) is approximately

$$v(\hat{t}_i) \approx \frac{2[1 + (n - 1)t_i]^2 (1 - t_i)^2 (sn - 1)}{s(s - 1)n^2 (n - 1)} \qquad [5]$$

with $E[\hat{t}_i] \simeq t_i$, and s and n, as before, the number of sires and progeny per sire. For any (fixed) set of parameters, the asymptotic distribution of the LR is a noncentral chi-square if the alternative hypothesis holds (11). For any replicate, the variance among estimated parameters, say ICC, had two components: one that is due to variance in true ICC, and one that is due to sampling variance about the true ICC. Therefore, a random effects model was assumed for the prediction of the power of the LR test. For a balanced one-way random effects model, the LR test is equivalent to an *F* test (9) and, therefore, asymptotically (for large within-group degrees of freedom) equivalent to a scaled central chi-square with (k – 1) degrees of freedom, where k is the number of groups. The scaling parameter depends on the ratio of the between- and within-group variance. Analogously, the variance of true parameters ("between" variance) and the approximate sampling variance of the estimated parameters ("within" variance) were used in the prediction of the power. The powers of tests 1 to 3 were predicted using

$$P(\alpha) = \int_{[\chi^2_\alpha(df)]/c}^{\infty} f(x)dx, \qquad [6]$$

with f(x) being the density of a central $\chi^2$ distribution with df degrees of freedom; $\chi^2_\alpha(df)$ is the 100(1 – α) percentage point for a central chi-square distribution.

The constant $c = [(v(\hat{\theta}_i|\theta_i) + v(\theta_i))/v(\hat{\theta}_i|\theta_i)]$ for hypotheses 2 and 3, $\theta_i = t_i$ for hypothesis 2, and $\theta_i = \sigma_i$ for hypothesis 3. For hypothesis 1, $c_1 = (c_2 + c_3)/2$; $c_2$ and $c_3$ are the constants for hypotheses 2 and 3, respectively.

### Balanced Cross-Classified Half-Sib Designs

If sires and herds (strata) are cross-classified, i.e., all sires have progeny in all herds, then the following questions arise. 1) What is the contribution of the additional information, i.e., that animals in different herds are related to one another, to the detection of heterogeneity of parameters? 2) What is the effect of assuming a hierarchical design when maximizing the likelihood when data were generated from a cross-classified design?

The implicit assumption in the latter question, that data from different herds were statistically independent of each other, was made for computational reasons by Swalve and Van Vleck (20), Van Vleck and Dong (22), Van Vleck et al. (23), and Visscher et al. (24) because relationships between animals in different herds were ignored in those studies. These questions were addressed again by using simulation. The following model was used to generate data consisting of mean square between sires within a stratum (MSB), mean square within sires within a stratum (MSW), and mean cross product for sires between strata (MCPB):

$$y_{ijl} = \alpha_i S_j + \beta_i e_{ijl} \qquad [7]$$

where $y_{ijl}$ is an observation on progeny l (l = 1,n) of sire j (j = 1,s) in stratum i (i = 1,k) with residual $e_{ijl}$, and $\alpha_i$ and $\beta_i$ are constants scaling the sire and residual variance. Therefore, the assumption is that additive genetic correlations between sire performances in different strata are unity and that a sire by herd interaction is the effect of scaling. Then, if **M** is a k × k matrix of MSB and MCPB between k strata, and if **W** is the diagonal matrix of MSW,

$$\begin{aligned} E[M_{ii}] &= \beta_i^2\sigma_w^2 + n\ \alpha_i^2\sigma_b^2 = \sigma_{wi}^2 + n\sigma_{bi}^2 \\ E[M_{im}] &= n\ \sigma_i\alpha_m\sigma_b^2 = n\ \sigma_{bi}\sigma_{bm} \\ E[W_i] &= \beta_i^2\sigma_w^2 = \sigma_{wi}^2, \end{aligned}$$

for strata i and m. The likelihood function was parameterized in terms of between- and within-sire components and was maximized conditional on the MSW being the MLE of the within-sire component for that stratum, i.e., $\text{MLE}(\sigma^2_{wi}) = W_i$. This was done for computational reasons [see Appendix and (12)]. The estimates of the variances were, therefore, only approximations of MLE. Although the parameterization in between- and within-sire variances is different from the one used in the previous section, the main interest is the power of detecting heterogeneous between-sire components, and this power is likely to be very similar to the power of detecting heterogeneous ICC. To verify this, a nested (hierarchical) design was simulated as in the previous section, but with parameterization of the likelihood function in between- and within-sire components. The effect of fixing the estimates of the within-sire components to the within mean squares is unlikely to have a great effect on the LR: even for the smallest design, the degrees of freedom for the within-sire components were as large as 270 (= 30 × (10 − 1)). Data were generated as follows for each replicate: for each herd, $t_i$ and $\sigma^2_i$ were sampled as in the previous section, and the scaling parameters were calculated as $\alpha^2_i = (t_i\sigma^2_i)/(t_0\sigma^2_0)$, $\beta^2_i = [(1 - t_i)\sigma^2_i]/[(1 - t_0)\sigma^2_0]$ with $t_0 = .1$ and $\sigma_0 = 1.0$. Sire progeny means and residuals were sampled from a normal distribution (using [7]), and MSW, MSB, and MCPB were accumulated. Simulations were performed for the same groups sizes as in the previous section. The number of replicates was 5000 and 1000 for group sizes of k = 10 and k = 25, respectively.

## RESULTS

### Balanced Nested Half-Sib Designs

Simulation results for small and medium group sizes for a balanced nested half-sib design are in Tables 1 and 2. The coefficients of variation, rather than the variances of the population parameters, were displayed to make comparisons between the powers for $t_i$ and $\sigma^2_i$. The design from Table 1 was chosen to give similar standard errors of the heritability ($h^2$ = 4t) estimates, as were obtained by Visscher et al. (24) using field data. For the parameters used in Table 1, the approximate standard error

TABLE 1. Observed ($O_i$) and predicted ($P_i$) powers (percentage) for likelihood ratio tests from a balanced half-sib design for 25 herds, 30 sires per herd, and 10 progeny per sire.[1,2,3,4]

| $CV(t_i)$[5] | $CV(\sigma^2_i)$[6] | $O_1$ | $P_1$ | $O_2$ | $P_2$ | $O_3$ | $P_3$ |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 5.5 | 5.0 | 5.5 | 5.0 | 4.5 | 5.0 |
| .1 | 0 | 7.1 | 6.4 | 8.2 | 7.1 | 4.4 | 5.0 |
| .1 | .1 | 87.4 | 84.1 | 7.3 | 7.1 | 89.7 | 90.8 |
| .2 | 0 | 13.9 | 11.8 | 16.6 | 15.7 | 4.5 | 5.0 |
| .2 | .2 | 100.0 | 100.0 | 14.8 | 15.7 | 100.0 | 100.0 |
| .3 | 0 | 30.1 | 24.9 | 37.0 | 35.6 | 4.3 | 5.0 |
| .3 | .3 | 100.0 | 100.0 | 38.4 | 35.6 | 100.0 | 100.0 |
| .4 | 0 | 51.0 | 47.4 | 62.6 | 62.6 | 4.6 | 5.0 |
| .4 | .4 | 100.0 | 100.0 | 63.3 | 62.6 | 100.0 | 100.0 |
| .5 | 0 | 70.6 | 72.4 | 80.3 | 84.1 | 4.9 | 5.0 |
| .5 | .5 | 100.0 | 100.0 | 82.6 | 84.1 | 100.0 | 100.0 |

[1] $SE(O_1)$ .0 to .7%; $SE(O_2)$ .4 to 1.2%; $SE(O_3)$ .0 to .5%. Empirical standard errors were calculated from 5000 replicates.

[2] Subscripts 1 to 3 for observed and predicted powers refer to different null hypotheses: 1 = both intraclass correlation (ICC) and phenotypic variances homogeneous; 2 = ICC homogeneous, allowing for heterogeneous phenotypic variances; 3 = phenotypic variances homogeneous, allowing for heterogeneous ICC.

[3] Mean ICC and phenotypic variance in the population are .1 and 1.0, respectively.

[4] Powers for $\alpha$ = 5%.

[5] Coefficient of variation of true individual herd ICC.

[6] Coefficient of variation of individual herd phenotypic variances.

TABLE 2. Observed ($O_i$) and predicted ($P_i$) powers (percentage) for likelihood ratio tests from a balanced half-sib design for 10 herds, 100 sires per herd, and 10 progeny per sire.[1,2,3,4]

| $CV(t_i)$[5] | $CV(\sigma_i^2)$[6] | $O_1$ | $P_1$ | $O_2$ | $P_2$ | $O_3$ | $P_3$ |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 4.8 | 5.0 | 5.5 | 5.0 | 5.2 | 5.0 |
| .1 | 0 | 8.4 | 8.2 | 10.2 | 10.0 | 5.3 | 5.0 |
| .1 | .1 | 95.2 | 96.9 | 10.3 | 10.0 | 96.3 | 96.3 |
| .2 | 0 | 25.3 | 22.7 | 29.7 | 31.1 | 5.0 | 5.0 |
| .2 | .2 | 100.0 | 100.0 | 30.9 | 31.1 | 100.0 | 100.0 |
| .3 | 0 | 55.5 | 51.5 | 63.1 | 62.4 | 4.8 | 5.0 |
| .3 | .3 | 100.0 | 100.0 | 64.4 | 62.4 | 100.0 | 100.0 |
| .4 | 0 | 78.9 | 79.1 | 84.2 | 84.1 | 4.6 | 5.0 |
| .4 | .4 | 100.0 | 100.0 | 82.6 | 84.1 | 100.0 | 100.0 |
| .5 | 0 | 89.8 | 93.4 | 92.6 | 94.0 | 4.7 | 5.0 |
| .5 | .5 | 100.0 | 100.0 | 92.4 | 94.0 | 100.0 | 100.0 |

[1] $SE(O_1)$ .0 to .6%; $SE(O_2)$ .3 to .8%; $SE(O_3)$ .0 to .4%. Empirical standard errors were calculated from 5000 replicates.

[2] Subscripts 1 to 3 for observed and predicted powers refer to different null hypotheses: 1 = both intraclass correlations (ICC) and phenotypic variances homogeneous; 2 = ICC homogeneous, allowing for heterogeneous phenotypic variances; 3 = phenotypic variances homogeneous, allowing for heterogeneous ICC.

[3] Mean ICC and phenotypic variance in the population are .1 and 1.0, respectively.

[4] Powers are for $\alpha = 5\%$.

[5] Coefficient of variation of true individual herd ICC.

[6] Coefficient of variation of individual herd phenotypic variances.

of the corresponding heritability estimate was .189 (from Equation [5]). The probability of rejecting the null hypothesis when it was true was very similar to the significance level for testing phenotypic variances and for testing heterogeneity of ICC. For the double homogeneity test, the LR test detected heterogeneity even when one of the parameters, in this case the phenotypic variance, was homogeneous (see columns pertaining to $O_1$ in Tables 1 and 2). Clearly, the power to detect heterogeneous ICC was very low compared with the power to detect differences in phenotypic variances. For example, if the $CV(t_i)$ in the pφpulation of herds was .3, which corresponds to a distribution of the heritability with mean of .40 and standard deviation of .12, then, in approximately 37% of repeated samples of 25 herd estimates, a difference in heritability would be detected. Table 2 appears to confirm that some of the (small) differences between observed and predicted powers in Table 1 were caused by small sample sizes. Again, the difference in power between LR tests for $t_i$ and $\sigma_i^2$ is striking. In general, simulation results agreed well with their predictions.

In Table 3, the predictions of the powers for large samples for two groups are shown. Such samples may be similar to estimating parameters from groups of herds that have been split according to the herd mean or herd variance. The standard error of the heritability is shown because results from studies investigating heterogeneity of variance in two or more groups (4, 10, 13) usually are reported in terms of differences between heritability estimates. Table 3 shows that, even for large sample sizes, moderate powers can be obtained using an LR test. For all sample sizes in Table 3, the predicted power of an LR test for detecting heterogeneity of phenotypic variances was 100% when taking the same range of coefficient of variation for the phenotypic variance as was used for the ICC.

### Balanced Cross-Classified Half-Sib Designs

Table 4 shows the results from simulating data from a balanced cross-classified design. Results are shown only for cases in which $CV(\sigma_i^2) = 0$, i.e., $CV(t_i) = CV(\sigma_{bi}^2/\sigma_i^2) = CV(\sigma_{bi}^2)$. Hence, between- and within-sire variances were heterogeneous, but their sum, the

TABLE 3. Predicted powers (for $\alpha = 5\%$) for detection of heterogeneous intraclass correlations (ICC) in two groups for likelihood ratio tests from various balanced half-sib designs, assuming that phenotypic variances are homogeneous.[1]

| | | | Power | | |
|---|---|---|---|---|---|
| s[2] | n[3] | SE($\hat{h}^2$)[4] | .1[5] | .2[5] | .3[5] |
| | | | | (%) | |
| 100 | 25 | .071 | 17 | 59 | 88 |
| | 50 | .061 | 22 | 72 | 94 |
| | 100 | .056 | 26 | 78 | 96 |
| 250 | 25 | .045 | 40 | 91 | 99 |
| | 50 | .039 | 52 | 96 | 100 |
| | 100 | .035 | 59 | 97 | 100 |
| 500 | 25 | .032 | 69 | 99 | 100 |
| | 50 | .027 | 80 | 99 | 100 |
| | 100 | .025 | 86 | 100 | 100 |
| 750 | 25 | .026 | 84 | 100 | 100 |
| | 50 | .022 | 91 | 100 | 100 |
| | 100 | .020 | 94 | 100 | 100 |
| 1000 | 25 | .022 | 91 | 100 | 100 |
| | 50 | .019 | 96 | 100 | 100 |
| | 100 | .018 | 97 | 100 | 100 |

[1] Mean ICC across the two groups is .1.

[2] Number of sires per group.

[3] Number of progeny per sire.

[4] Using $\hat{h}^2 = 4\hat{t}$ and Equation [5]

$$v(\hat{t}_i) \approx \frac{2[1 + (n - 1)t_i]^2 (1 - t_i)^2 (sn - 1)}{s(s - 1)n^2 (n - 1)} .$$

[5] CV($t_i$) = Coefficient of variation of true ICC between the two groups.

phenotypic variance, was the same for all herds. The first columns for each of the two population designs, i.e., columns $O_4$, can be directly compared with columns $O_2$ from Tables 1 and 2. Clearly, the powers for detecting heterogeneous sire components and ICC are similar. The second column of observed powers in Table 4 shows the effect of assuming the incorrect model for calculating the LR. The loss in power occurs because part of the information about the covariance structure of the MSB is not taken into account in the calculation of the maximum likelihood. Note that the estimates of the between- and within-sire components both for the unrestricted model (different between- and within-sire components for each stratum) and for the $H_0$ hypothesis are unbiased (conditional on the ANOVA estimates for the between-sire variance being positive), because the expectations of the mean squares in the usual ANOVA are not changed; ignoring the MCPB simply means that the variance of the estimates is increased. The estimated Type 1 errors for both designs were less than 1% if an incorrect model was assumed, at a nominal significance level of 5%. If the null hypothesis was false, the probability of rejecting it, i.e., power, was also low (see columns $O_5$).

The final column in Table 4 indicates the gain of using MCPB for the assumed model to detect heterogeneous variance components. The power was increased substantially, in particular for the range of CV($t_i$) of .2 to .3. In absolute terms, the power was still small for design 1 (25 strata, 30 sires, 10 progeny per sire): if the coefficient of variation of the between-sire variance was .30 in the population, this heterogeneity would be picked up in approximately 62% of samples. For CV($t_i$) = .1, the power for the nested design (8.2%) was found to be larger than the power for the cross-classified design (7.7%) for the design with 25 herds, although a larger power was expected for the cross-classified design. This may be explained by sampling (standard errors of mean powers were .3 and .5, respectively) and by departures from normality for small sample estimates. The estimated Type 1 error for the nested design (column $O_4$) was 6.6%, at a nominal significance level of 5%, whereas the estimated Type 1 error for the cross-classified design was 5.3%.

## DISCUSSION

The analytical and simulation results show clearly that the power of an LR test for detecting heterogeneous ICC (or heritabilities) is very low for the range of standard errors of heritability estimates to be expected from individual herd data in most countries. Visscher et al. (24) used 6 yr of first lactation data from 26 large pedigree herds in England and Wales and obtained standard errors of heritability estimates of approximately .19. Van Vleck and Dong (22), using 300 to 400 first lactation records per herd, estimated the standard errors of their heritability estimates to be approximately .15. The United Kingdom has the largest average herd size in Europe, so sampling variances of individual herd estimates would be larger in other countries in Europe.

TABLE 4. Observed ($O_i$) powers and standard errors (percentage), for $\alpha = 5\%$, in detecting heterogeneous sire variances for likelihood ratio (LR) tests from balanced nested and cross-classified half-sib designs when phenotypic variances are homogeneous.[1,2,7]

| Design 1, k[3] = 25, s[4] = 30, n[5] = 10 | | | | | | | Design 2, k[3] = 10, s[4] = 100, n[5] = 10 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CV($\sigma_b^2$)[6] | $O_4$ | | $O_5$ | | $O_6$ | | $O_4$ | | $O_5$ | | $O_6$ | |
| | | SE | | SE | | SE | | SE | | SE | | SE |
| 0 | 6.6 | .3 | .4 | .2 | 5.3 | .3 | 5.7 | .5 | .9 | .1 | 5.1 | .3 |
| .1 | 8.2 | .3 | .5 | .2 | 7.7 | .5 | 10.0 | .5 | 2.5 | .3 | 14.1 | .3 |
| .2 | 16.5 | .6 | 2.1 | .6 | 28.6 | 1.1 | 29.4 | .9 | 15.9 | .4 | 49.9 | .6 |
| .3 | 35.6 | .9 | 12.2 | .9 | 61.6 | 1.5 | 58.5 | .7 | 49.5 | .5 | 82.8 | .7 |
| .4 | 58.6 | .4 | 32.2 | 1.8 | 86.5 | 1.1 | 80.8 | .4 | 74.0 | .7 | 94.7 | .2 |
| .5 | 76.5 | .6 | 57.8 | 2.0 | 96.4 | .6 | 91.3 | .4 | 88.0 | .7 | 98.5 | .2 |

[1] All LR are conditional on D = W.

[2] Subscripts for observed powers 4 to 6 refer to the following data structures and null hypotheses $H_0$: 4 = data from nested design, $H_0$ = homogeneous sire variances; 5 = data from cross-classified design, but ignoring mean cross product for sires between strata, $H_0$ = homogeneous sire variances; 6 = data from cross-classified design, $H_0$ = homogeneous sire variances.

[3] Number of herds.

[4] Number of sires per herd.

[5] Number of progeny per sire.

[6] Coefficient of variation for sire variance across herds.

[7] Empirical standard errors were calculated from 1000 replicates (design 1) and 5000 replicates (design 2).

Using more records per herd seems obvious but may give additional problems of heterogeneity of variance between herd-years and between lactations, if the use of later lactations is considered.

Therefore, the conclusion of Visscher et al. (24), that heritability estimates were fairly homogeneous and that phenotypic variances differed between herds, is not surprising given the low power of the statistical test. However, before using an IAM-BLUP evaluation, a decision should be made with regards to the correct covariance structure of the data. Given the lack of power in detecting any differences in heritabilities between herds, it seems logical to assume that heritabilities are homogeneous. Records can then be scaled according to a (regressed) estimate of the within-herd phenotypic variances if those variances are found to be heterogeneous. A Bayesian interpretation for assuming homogeneous heritability for practical purposes is that the individual herd estimates should be regressed to an overall heritability estimate (a prior for the mean of the distribution of the heritability), and, because the sampling variances of the individual estimates are large, the regressed estimates would be very similar (homogeneous).

Shaw (17) used stimulation to investigate powers to detect differences in additive genetic (co)variances between two populations. Using a balanced hierarchical design of dams within sires and an LR significance test, Shaw (17) found low powers to detect differences between the two populations. For example, when the additive genetic variance in the populations differed by a factor of 2.5, the power was approximately 50% for a design of 100 sires, 3 dams per sire, and 3 progeny per dam for each population.

Foulley et al. (6) presented a general framework to test for sources (e.g., herds or sires) causing heterogeneity of residual variance and presented an example to illustrate the generality of their test. However, the test failed to detect heterogeneity of residual variance caused by sires, and it may be argued that the power of the presented hypothesis test, essentially an LR test, for detecting heterogeneity of sire variances (whether caused by herds or sires) is likely to be low in most practical situations. San Cristobal et al. (16) questioned the robustness of their or any LR test to departures from normality, but results from the half-sib designs seem to suggest that, for relatively small samples, the lack of statistical power is of greater practical importance than violations of normality assumptions.

The power for large samples approaches unity rapidly (Table 3), although differences in t (heritability) may not be detected for two herd-groups with 100 to 200 sires represented. For example, Hill et al. (10) estimated parameters in two (high and low) groups, each with 762 sires and approximately 11 effective daughters per sire. Using the prediction formula [6], with t = .0625 ($h^2$ = .25) and $\alpha$ = 5%, repeated samples of 2 herd groups from the total population would give a power of 13, 32, 47, 58, and 65% for CV($h^2$) = .1, .2, ..., .5, respectively. These relatively low powers are confirmed by performing a simple significance test on the difference of the estimates in the high and low group, assuming that heritability estimates are normally distributed. Although the sign of the difference is consistent (high mean and high variance groups showed higher heritabilities), the test statistic is not significant at the 5% level.

Using information between herds or strata may increase the power of the LR test, but simplified models are necessary, for computational reasons, to make calculation of likelihoods under various hypotheses feasible. If, for example, in the cross-classified design, the assumption about scaling was not made, the number of between-sire parameters to be estimated would increase from k to k (k + 1)/2.

## CONCLUSION

The power of detecting heterogeneous heritabilities or (additive) genetic variances between herds using field data is expected to be small, but it is relatively easy to detect differences in total phenotypic variances.

## ACKNOWLEDGMENTS

Financial support from the Milk Marketing Boards in the United Kingdom is appreciated. I wish to thank Bill Hill, whose questioning prompted this study, for helpful comments. I thank Robin Thompson for suggesting investigation of the cross-classified design, for suggesting an appropriate computing algorithm, and for useful comments. I thank both referees for their helpful comments.

## REFERENCES

1 Bartlett, M. S. 1937. Properties of sufficiency and statistical tests. Proc. R. Soc. Lond. A 160:268.

2 Boldman, K. G., and A. E. Freeman. 1990. Adjustment for heterogeneity of variances by herd production level in dairy cow and sire evaluation. J. Dairy Sci. 73:503.

3 Brotherstone, S., and W. G. Hill. 1986. Heterogeneity of variance amongst herds for milk production. Anim. Prod. 42:297.

4 Dong, M. C., and I. L. Mao. 1990. Heterogeneity of (co)variance and heritability in different levels of intraherd milk production variance and herd average. J. Dairy Sci. 73:843.

5 Fisher, R. A. 1921. On the "probable error" of a coefficient of correlation deduced from a small sample. Metron 1 part 4:1.

6 Foulley, J. L., D. Gianola, M. San Cristobal, and S. Im. 1990. A method for assessing extent and sources of heterogeneity of residual variances in mixed linear models. J. Dairy Sci. 73:1612.

7 Reference deleted in proof.

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

9 Herbach, L. H. 1959. Properties of model II-type analysis of variance tests, A: Optimum nature of F-test for model II in the balanced case. Ann. Math. Stat. 30: 939.

10 Hill, W. G., M. R. Edwards, M.-K.A. Ahmed, and R. Thompson. 1983. Heritability of milk yield and composition at different levels and variability of production. Anim. Prod. 36:59.

11 Kendall, M. G., and A. Stuart. 1973. Page 230 *in* The advanced theory of statistics. 3rd ed., Vol. 2. Inference and relationship. Griffin, London, Engl.

12 Lawley, D. N., and A. E. Maxwell. 1971. Factor analysis as a statistical method. 2nd ed. Butterworth & Co., London, Engl.

13 Lofgren, D. L., W. E. Vinson, R. E. Pearson, and R. L. Powell. 1985. Heritability of milk yield at different herd means and variances for production. J. Dairy Sci. 68:2737.

14 Mirande, S. L., and L. D. Van Vleck. 1985. Trends in genetic and phenotypic variances for milk production. J. Dairy Sci. 68:2278.

15 Patterson, H. D., and R. Thompson. 1971. Recovery of inter block-information when block sizes are unequal. Biometrika 58:545.

16 San Cristobal, M., J. L. Foulley, and D. Gianola. 1990. Inference about heterogeneous U-components of variance in mixed linear models. 41st Eur. Assoc. Anim. Prod. Mtg., Toulouse, France.

17 Shaw, R. G. 1991. The comparison of quantitative genetic parameters between populations. Evolution 45: 143.

18 Short, T. H., R. W. Blake, R. L. Quaas, and L. D. Van Vleck. 1990. Heterogeneous within-herd variance. 1. Genetic parameters for first and second lactation milk yields of grade Holstein cows. J. Dairy Sci. 73:3312.

19 Smith, S. P., and H.-U. Graser. 1986. Estimating variance components in a class of mixed models by restricted maximum likelihood. J. Dairy Sci. 69:1156.

20 Swalve, H., and L. D. Van Vleck. 1987. Estimation of genetic (co)variances for milk yield in first three

lactations using an animal model and restricted maximum likelihood. J. Dairy Sci. 70:842.

21 Thompson, R., and K. Meyer. 1986. Estimation of variance components: what is missing in the EM algorithm? J. Stat. Comput. Simul. 24:215.

22 Van Vleck, L. D., and M. C. Dong. 1988. Genetic (co)variances for milk, fat, and protein yield in Holsteins using an animal model. J. Dairy Sci. 71:3040.

23 Van Vleck, L. D., M. C. Dong, and G. R. Wiggans. 1988. Genetic (co)variances for milk and fat yield in California, New York, and Wisconsin for an animal model by restricted maximum likelihood. J. Dairy Sci. 71:3053.

24 Visscher, P. M., R. Thompson, and W. G. Hill. 1991. Estimation of genetic and environmental variances for fat yield in individual herds and an investigation into heterogeneity of variance between herds. Livest. Prod. Sci. 28:273.

25 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

### Estimation of Variances in Balanced Cross-Classified Half-Sib Designs

The following algorithm was suggested by R. Thompson.

Assume that a matrix M of MSB and MCPB and a diagonal matrix W of MSW are observed from k herds. Each of the s sires has n progeny in each herd. For herd i, the between- and within-sire variances are $\sigma^2_{bi}$ and $\sigma^2_{wi}$, respectively. For the "full" model, it is further assumed that

$$E[\mathbf{M}] = \mathbf{V} = \mathbf{LL'} + \mathbf{D}, \qquad [A1]$$

where **L** is a vector of length k with elements $L_i = n\,\sigma_{bi}$, and **D** is a diagonal matrix of order k with $D_i = \sigma^2_{wi}$.

Then the residual likelihood is

$$-2L_u(\mathbf{M}, \mathbf{W}|\mathbf{V}) = (s - 1)[\log|\mathbf{V}| + \mathrm{tr}(\mathbf{MV}^{-1})] + s(n - 1)[\log|\mathbf{D}| + \mathrm{tr}(\mathbf{WD}^{-1})]. \qquad [A2]$$

Conditional on **D** = **W**, and ignoring the second part of the likelihood pertaining to **D**, the maximum likelihood can be written as

$$-2ML_u(\mathbf{M}|\mathbf{V}, \mathbf{D} = \mathbf{W}) = (s - 1)[\log(\theta_1) + \sum_{i=2}^{k} \theta_i + \sum_{i=1}^{k} \log(W_i)] \qquad [A3]$$

where $\theta_i$ are the eigenvalues of $\mathbf{M}^* = \mathbf{D}^{-\frac{1}{2}}\mathbf{M}\,\mathbf{D}^{-\frac{1}{2}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{M}\,\mathbf{W}^{-\frac{1}{2}}$ and $\theta_1$ is the largest eigenvalue of **M***.

Hence, conditional on **D** = **W**, no iterative procedure is required to calculate the maximum likelihood for the full model. Unless the number of herds is very large, calculating the eigenvalues for a symmetric k × k matrix is computationally relatively easy. The algorithm is similar to a commonly used algorithm in factor analysis; the analogy is to regard sires as the only "factor" in the analysis explaining the data [see e.g., (11)].

Computation of the maximum likelihood for the alternative hypothesis, that all sire variances are the same, again assuming **D** = **W**, involves computing the MLE of the overall sire variance. Let the overall sire variance be $\sigma^2_{b0}$. It can be shown that for this model the MLE of $\sigma^2_{b0}$ has an explicit solution, which is

$$\mathrm{ML}(\hat{\sigma}^2_{b0}) = [(\mathbf{1}'\mathbf{D}^{-1}\mathbf{M}\mathbf{D}^{-1}\mathbf{1}) - (\mathbf{1}'\mathbf{D}^{-1}\mathbf{1})]/[n(\mathbf{1}'\mathbf{D}^{-1}\mathbf{1})^2]$$

where $\mathbf{1}'$ is a row vector of length k with all elements unity.

If data from different herds are assumed to be independent, computations of the maximum likelihood requires solving a cubic equation in $\sigma^2_{b0}$. The MLE of the common sire variance then satisfies, conditional on $\mathbf{D} = \mathbf{W}$,

$$\sum^{k} 1/(\sigma^2_{wi} + n\sigma^2_{b0}) = \sum^{k} M_i/(\sigma^2_{wi} + n\sigma^2_{b0})^2.$$

This is relatively straightforward to solve.