Estimating Parameters of a Random Regression Test Day Model for First Three Lactation Milk Production Traits Using the Covariance Function Approach

Z. Liu, F. Reinhardt and R. Reents
United Datasystems for Animal Production (VIT), Heideweg 1, D-27280 Verden, Germany

Introduction

Estimating (co)variance parameters of a random regression model for test day yields of dairy cattle has been proven to be a challenging task due to the complexity of (co)variance structure of the test day yields. Two common approaches have been practised so far for the estimation of parameters: direct random regression model approach and indirect covariance functions (CF) approach. Animal genetic and permanent environmental effects are modelled in the direct approach with a lactation curve function and (co)variance components of random regression coefficients (RRC) for both effects are estimated jointly with other effects in the model. By contrast, in the indirect CF approach (co)variance parameters of genetic and residual effects are estimated using a multivariate model that treats test day yields of different lactation stages as genetically distinct traits, and then CF are fitted to the estimated (co)variance matrices from the first step to obtain (co)variances of RRC for additive genetic as well as permanent environmental effects. The objective of this study was to estimate (co)variance components of RRC of a random regression test day model for first three lactation test day yields via the CF approach.

Material and Methods

A multi-trait sire model was applied to the first three lactation test day yields for each of the three biological traits, milk, fat and protein yields, to estimate (co)variance components via the CF approach:

$$y_{ijklmn} = \mu_{lm} + HTD_{il} + \sum_{p=1}^{5} \beta_{jlp} v_{pd} + s_{klm} + e_{ijklmn} \qquad [1]$$

where

y_{ijklmn} is test day yield at lactation stage *m* of lactation *l* of cow *n*,

μ_{lm} is general mean for lactation stage *m* of lactation *l*,

HTD_{il} is the *i*-th herd-test-date effect of lactation *l*,

v_{pd} is the *p*-th parameter of Ali-Schaeffer function for days in milk (DIM) *d*,

β_{jlp} is the *p*-th fixed regression coefficient for lactation *l* specific to subclass *j*,

s_{klm} is additive genetic effect of sire *k* for lactation stage *m* of lactation *l*, and

e_{ijklmn} is the residual effect

For each of the first three lactations six lactation stages are defined based on DIM: 5-50, 51-105, 106-160, 161-215, 216-259, 260-305. Test day yields from different lactation stages are treated as genetically distinct traits in model 1.

Raw data from February 2000 German Holstein genetic evaluation were selected based on the following criteria: *HTD* classes with at least five records, supervised monthly testing with two times milkings only, and calving years for first three lactations no earlier than 1993, 1994 and 1995 respectively. In case of duplicate test day records within a lactation stage, one record was randomly chosen. Only completed lactations were kept for estimating parameters. Sires with fewer than 30 daughters were discarded to achieve a better data structure. The original pedigree file from the routine genetic evaluation was used to extract pedigree information for all ancestors of cow sires. Table 1 shows the structures of the final test day data set and sire pedigree file used in parameter estimation. For each of the three lactations, 420 fixed lactation curves were fitted to data based on three calving seasons, five classes of age at calving, four breed-region classes and seven calving interval classes.

Table 1. Description of the final data set and sire pedigree file for parameter estimation

Factors	Cows	Sires of cow	Test day records in total	HTD of all lactations	Fixed lactation curves in total	Animals in sire pedigree file
Size	1,590,592	5,042	15,605,538	2,986,200	1,260	9,956

An iterative approach to estimating (co)variance components

Considering the fact that there are more equations of the fixed effects than random sire effects in model 1, a so-called iterative two-step approach (Gengler et al. 1999) was implemented in order to estimate the parameters efficiently:

Step 1. Estimating the fixed effects of HTD and lactation curves using ordinary least squares method with model 2:

$$y_{ijklmn} - (\hat{\mu}_{lm} + \hat{s}_{klm}) = HTD_{il} + \sum_{p=1}^{5} \beta_{jlp} v_{pd} + \varepsilon_{ijklmn} \quad [2]$$

where ε_{ijklmn} is residual effect and $\hat{s}_{klm}$ is sire estimated breeding value (EBV) from the previous round of iteration. For the first round of iteration sire EBVs from the routine evaluation with a fixed regression test day model (Reents et al., 1998) were used as starting values.

Step 2. Estimating (co)variance components of sire effects via restricted maximum likelihood using test day records adjusted for the fixed effects above:

$$y_{ijklmn} - (\hat{HTD}_{il} + \sum_{p=1}^{5} \hat{\beta}_{jlp} v_{pd}) = \mu_{lm} + s_{klm} + \xi_{ijklmn} \quad [3]$$

where ξ_{ijklmn} is residual effect. Variance components were estimated via VCE (Neumaier and Groeneveld 1998), in which analytical gradients of likelihood function are explicitly calculated instead of the approximation by finite differences and maximisation of the likelihood is done by quasi-Newton algorithm based on exact first derivatives. Mixed model equations are set up in memory using sparse matrix storage technique and solving the equations is done on the basis of Cholesky factorisation. Due to the large number of components to be estimated for three lactations, the estimation task was partitioned into seven 9-trait analyses to obtain parameter estimates of all 18 traits. The two steps are repeated until all (co)variance components and sire EBVs are converged. After the iteration process has been completed, simple averages of the (co)variance estimates from the seven parallel runs were calculated for later derivation of (co)variances of RRC. Sampling variances of the (co)variance component estimates were obtained from VCE as well.

Deriving (co)variances of RRC using Legendre polynomials of order three: A total of eight mathematical functions: Wilmink function, Ali-Schaeffer function, mixed log function, Legendre polynomials of order 2 to 5, and a quadratic function of DIM, were selected to derive and compare (co)variances of RRC using the estimated (co)variance matrices (Liu et al., unpublished data). Based on the comparison results, the third-order normalised orthogonal Legendre polynomial was chosen to derive (co)variances of RRC for first three lactation test day yields:

$$a_1 + a_2\sqrt{3}z + a_3\frac{\sqrt{5}}{2}(3z^2 - 1) \ ,$$

where *a*'s represent RRC, $z = [2(x - \min) - (\max - \min)] / (\max - \min)$ with min and max being minimum and maximum DIM values, and *x* denotes DIM.

Additive genetic ($\mathbf{G}_s$) and residual ($\mathbf{R}_s$) (co)variance matrices of the sire model 1 were converted to an animal model basis: $\mathbf{G} = 4\mathbf{G}_s$, $\mathbf{R} = \mathbf{R}_s - 3\mathbf{G}_s$. There have been so far two methods, generalised least square inverse

(Tijani et al., 1999) and expectation-maximisation algorithm (Mäntysaari, 1999), developed in animal breeding to derive (co)variances of RRC based on the estimated (co)variance matrices, **G** and **R**. Compared to Kirkpatrick et al.'s approach (1994), both methods ignore the fact that the (co)variance estimates are usually associated with different standard errors. For instance, the (co)variances of first lactation are more precisely estimated than those of later lactations as a result of more test day records in first lactation. In this study Kirkpatrick et al.'s approach as well as the generalised least squares inverse approach were modified to derive (co)variances of RRC for both additive genetic and permanent environmental effects.

In the modified generalised least squares inverse approach, denoted as *Tm*, covariance matrix of additive genetic RRC, $\mathbf{K_G}$, is computed in the same way as in Tijani et al. (1999):

$$\mathbf{K_G} = (\mathbf{Z'Z})^{-1}\mathbf{Z'GZ}(\mathbf{Z'Z})^{-1} \qquad [4]$$

where **Z** is the incidence matrix for additive genetic RRC. An iterative procedure to separate time-dependent permanent environmental effects from time-independent error effects has been developed under the assumption that for a biological trait error effects are independently distributed as $N(0,\mathbf{E})$ with a diagonal block $\mathbf{E}_j = \mathbf{I}\sigma^2_{e(j)}$ for lactation *j*. Note that this iterative procedure is not restricted to the assumption of a constant error variance throughout the course of lactation and it can also model heterogeneous error variances (Jamrozik and Schaeffer, 1997), e.g. via a link function approach (Jaffrezic et al., 2000). Prior to the iteration, starting values are assigned to (co)variance matrix of permanent environmental RRC that must be positive definite ($\mathbf{K_P^{[0]}}$) and (co)variance matrix of permanent environmental effects is computed with $\mathbf{P}^{[0]} = \mathbf{ZK_P^{[0]}Z'}$. At iteration round *i* average permanent environmental variance $\bar{\sigma}^{2[i]}_{p(j)}$ is obtained for each of the six lactation stages of lactation *j* and then the average is subtracted from corresponding diagonal element of **R**, σ^2_R. Error variance $\sigma^{2[i]}_{e(j)}$ is set equal to the average of the differences (σ^2_R - $\bar{\sigma}^{2[i]}_{p(j)}$) across all lactation stages of lactation *j* and error variance matrix is updated as $\mathbf{E}^{[i]}_j = \mathbf{I}\sigma^{2[i]}_{e(j)}$. (Co)variance matrices of permanent environmental effects and permanent environmental RRC for the next round are updated with $\mathbf{P}^{[i+1]} = \mathbf{R} - \mathbf{E}^{[i]}$ and $\mathbf{K_P^{[i+1]}} = (\mathbf{Z'Z})^{-1}\mathbf{Z'P}^{[i+1]}\mathbf{Z}\,(\mathbf{Z'Z})^{-1}$. The above steps are repeated until $\mathbf{K_P}$ and **E** are converged. This *Tm* approach appears to guarantee that the estimated matrices of RRC are positive definite.

The extended weighted least squares method, denoted as *Km*, estimates (co)variances of additive genetic RRC in the same way as Kirkpatrick et al. (1994) proposed with the following weighted least squares equations:

$$\tilde{\mathbf{c}} = (\mathbf{X}_s'\hat{\mathbf{V}}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s\hat{\mathbf{V}}^{-1}\hat{\mathbf{g}} \qquad [5]$$

where $\tilde{\mathbf{c}}$ vector contains elements of $\mathbf{K_G}$ above diagonal, $\hat{\mathbf{g}}$ vector contains elements of **G** above diagonal, **V** matrix has diagonal elements being sampling variances of all elements in $\hat{\mathbf{g}}$, and $\mathbf{X}_s$ is determined by the Legendre polynomial function (Kirkpatrick et al. 1994). For separating permanent environmental (co)variances from the residual (co)variance **R**, the same iterative approach as in the method *Tm* is implemented, except that the (co)variances of RRC for permanent environmental effects are estimated using the weighted least squares method as for additive genetic effects in the method *Km*. Note that the method *Km* is based on the similar idea as Kirkpatrick et al.'s approach *Extrapolating to the Diagonal*.

Results and Discussion

Fortran 90 programs and Unix shell scripts were developed for estimating the parameters of the multiple trait sire model. The computation was conducted on a HP9000 K460 computer running HP-UX 11 and a

Pentium III PC running Linux. Four iterative steps were required to get both (co)variance estimates and sire EBVs converged, with mean relative differences between iteration steps being less than 1%. To derive (co)variances of RRC based on the (co)variance estimates of the sire model, Maple V programs were written for Microsoft Windows system. The whole estimation took considerable CPU and memory resources. For the derivation of (co)variance matrix of RRC for permanent environmental effects, $\mathbf{K}_P$, 25 iteration steps were needed to reach convergence using either method *Tm* or *Km***.** Since both methods *Tm* and *Km* result in similar (co)variances of RRC, indicating the high accuracy of the (co)variance estimates of the sire model, the derived (co)variance matrices of RRC with the method *Km* were then chosen for subsequent analyses.

Derived heritability values on daily basis: Figures 1 to 3 show daily heritability values, derived based on the (co)variances of RRC, of first three lactation test day milk, fat and protein yields, respectively. For test day milk yields, daily heritability ranges from 0.12 at DIM 5 for third lactation to 0.36 at DIM 180 for first lactation. It is obvious that the beginning and end of lactation have lower heritability than the middle part. Both the level and the pattern of daily heritabilities are as normally expected. Low heritability values were not found here using the CF approach as in Mäntysaari (1999), Strabel and Misztal (1999), and Tijani et al. (1999). Later lactations have clearly lower heritabilties than the first one, and no evident difference in heritability was observed between second and third lactations. Among the three production traits, milk yield has the highest heritability.

Figure 4 shows the ratios of genetic standard deviations of two lactations for test day milk yields. The ratio of second or third lactation to first lactation is near one from early lactation stages up to DIM 200, and then it increases rapidly towards the end of lactation. By contrast, the ratio between second and third lactations stays fairly flat throughout the course of lactation.

Figure 1. Heritability values of first three lactation test day milk yields

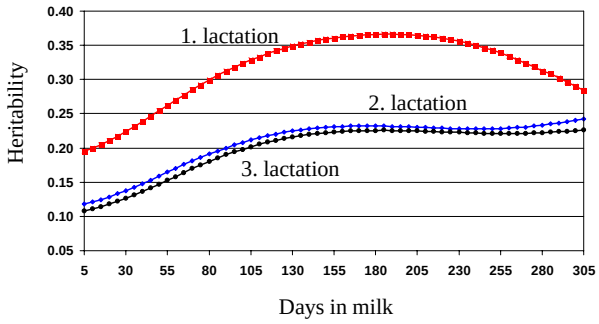
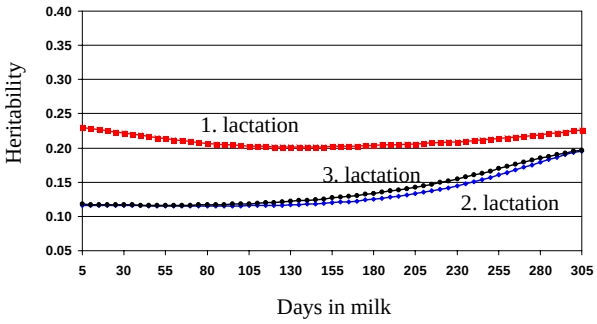


Figure 2. Heritability values of first three lactation test day fat yields



Derived genetic correlation structure: DIM 30 at the beginning, DIM 150 in the middle, and DIM 250 at the end of lactation were chosen to represent the genetic correlation structure of test day yields and the results are shown in Figures 5 and 6. Genetic correlation between two ends of lactation, e.g. between DIM 30 and 305 or between DIM 5 and 250, is about 0.5 or 0.45 for first lactation milk yield, which agrees with other investigations (Mäntysaari, 1999; Strabel and Misztal, 1999; Tijani et al., 1999). Daily yields of second lactation are less correlated than daily yields of first lactation. Fat and protein yields show a similar genetic correlation structure as milk yields. The use of the biological lactation curves, Wilmink, Ali-Schaeffer and mixed log functions, to derive (co)variances of RRC, resulted in negative genetic correlations between early and late lactation stages, whereas the general purpose functions, Legendre polynomials and quadratic function of DIM, can reproduce the original structure of the genetic (co)variance matrix **G** (Liu et al., unpublished data).

Figure 3. Heritability values of first three lactation test day protein yields

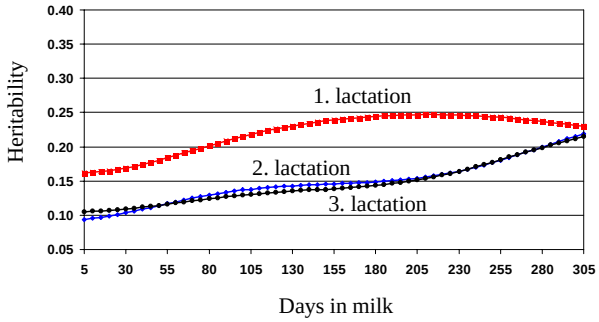


Figure 4. Ratio of genetic standard deviations between two of the first three lactations for test day milk yields

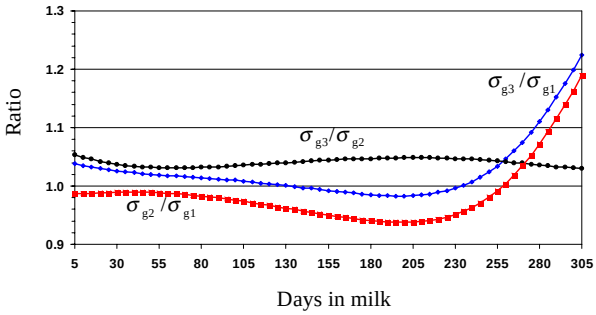


Figure 7 presents genetic correlations between the same DIM of two lactations for test day milk yield. Between first and second lactations the genetic correlation ranges from 0.73 at DIM 305 to 0.86 at DIM 155. A very similar genetic correlation curve was observed between first and third lactations. It can be seen that the middle stages of lactation are more highly correlated between lactations than the two ends of lactation. The genetic correlation of the same DIM between second and third lactation is quite high, above 0.95, which indicates that the second and third lactations are genetically very similar. Correlation structure for permanent environmental effects was also derived in the same way. In general, permanent environmental effects at different DIM are less correlated than additive genetic effects.

Figure 5. Genetic correlations between a given DIM and the remaining part of lactation for first lactation test day milk yields

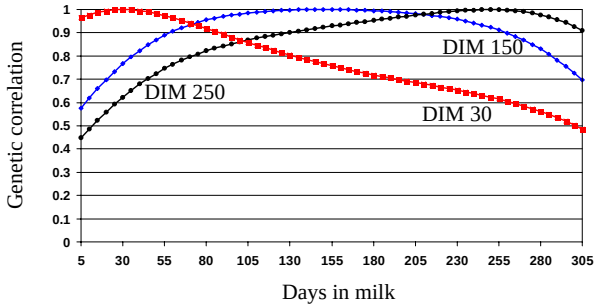
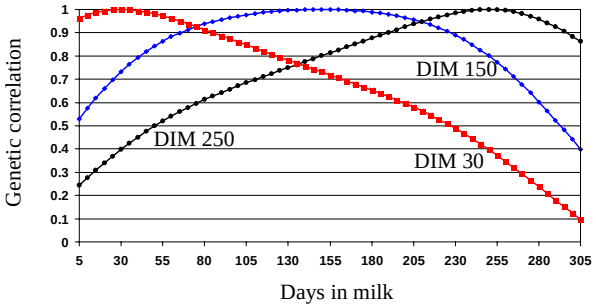


Figure 6. Genetic correlations between a given DIM and the remaining part of lactation for second lactation test day milk yields



Derived phenotypic correlation structure: Phenotypic correlation between any two DIM of the same lactation or of two different lactations was calculated using the (co)variance matrices of RRC for additive genetic ($\mathbf{K_G}$) and permanent environmental ($\mathbf{K_P}$) effects plus error variances ($\mathbf{E}$). Discrepancies between the derived and observed phenotypic correlation structures may indicate erroneous (co)variance estimates of RRC. Figure 8 displays the derived phenotypic correlations between DIM 30, 150 or 250 and the remaining part of lactation for first lactation test day milk yield. It should be noted that the phenotypic correlation at the selected DIM 30, 150 or 250 corresponds to the repeatability value at the same DIM. The estimated phenotypic correlations between the two ends of lactation, e.g. 0.27 between DIM 30 and 305 or 0.24 between DIM 5 and 250, agree with the observed phenotypic correlation.

The heritability estimates are consistent with those using a fixed regression test day model (Reents et al., 1995) and also with those using a random regression test day model (White et al., 1999). Compared to first lactation, second and third lactations have

much lower heritabilities for all three production traits, that was also found in a random regression model (Strabel and Mistzal, 1999) and a 305-day lactation model (Visscher and Thompson, 1992). By contrast, a reverse trend in heritability values was found in Jamrozik et al. (1997). Genetic correlation is, on average, 0.81 between first and second lactations for milk yield, and a similar estimate was obtained by Strabel and Misztal (1999) as well. However, lower genetic correlation was reported in Canadian test day model (Interbull Bulletin, No 24).

Compared to the Legendre polynomials and quadratic function of DIM, the biological lactation curve functions, Wilmink, Ali-Schaeffer and mixed log functions, show an inability to model the association between yields in early and late lactation stages, resulting in negative genetic correlation between two ends of lactation. This problem was reported by Brotherstone et al. (1999) too. In addition to the negative genetic correlation, additive genetic RRC derived using the three biological lactation curves are more highly correlated than those using the general purpose functions, with the highest correlations obtained from Ali-Schaeffer function. This means that more rounds of iteration would be needed for test day model genetic evaluation when the biological lactation curve functions were used to describe the (co)variance structure of test day yields than the general purpose lactation curve functions.

Figure 7. Genetic correlations between the same DIM of two of first three lactations for test day milk yields

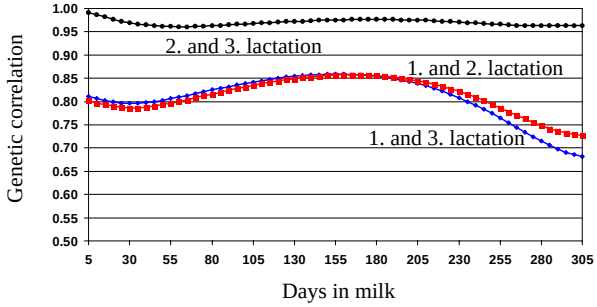
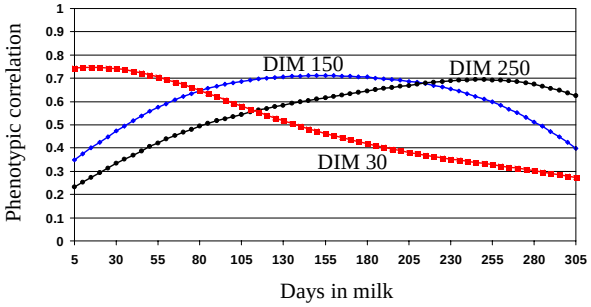


Figure 8. Phenotypic correlation between a given DIM and the remaining part of lactation for first lactation test day milk yields



Summary

A multiple trait sire model was applied to a very large test day data set for estimating (co)variance components of different lactation stages using an iterative two-step procedure. A third-order Legendre polynomial was subsequently fitted to the estimated (co)variances to derive (co)variances of RRC for both additive genetic and permanent environment effects. Tijani et al's generalised least square inverse approach and Kirkpatrick et al.'s weighted least squares approach have been modified to separate (co)variances of permanent environmental effects from error effects. Heritability estimates on daily basis are not as low as in some studies where the CF approach was also implemented, and the heritability estimates are consistent with several studies using both fixed and random regression test day models. Genetic correlations between any two DIM of the same lactation lie within a biologically acceptable range. Also the genetic correlations between the same DIM of two lactations are close to estimates in some previous investigations. The derived phenotypic correlations correspond well with observed ones. It was found in this study that the widely used biological lactation curve functions, Wilmink, Ali-Schaeffer and mixed log functions, can not model the (co)variance structure of test day yields correctly, whereas the general purpose functions, Legendre polynomials and quadratic function of DIM, are able to reproduce the original structure of (co)variances of test day yields. The presented estimation procedure has been shown to be feasible to analyse very large test day data set and can give highly accurate (co)variance components.

Literature Cited

Brotherstone, S., White, I.M.S. & Meyer, K. 1999. Genetic modelling of daily milk yield using orthogonal polynomials and parametric curves. In press.

Gengler, N., Tijani, A. Wiggans, G.R., Van Tassel, C.P. & Philpot, J.C. 1999. Estimation of (co)variances of test day yields for first lactation Holsteins in the United States. *J. Dairy Sci. 82*, (Aug). Online.

Jaffrezic, F., White, I.M.S., Thompson, R. & Hill, W.G. 2000. A link function approach to model heterogeneity of residual variances over time in lactation curve analyses. *J. Dairy Sci. 83*, 1089-1093.

Jamrozik, J. & Schaeffer, L.R. 1997. Estimates of genetic parameters for a test day model with random regression for yield traits of first lactation Holsteins. *J. Dairy Sci. 80*, 762-770.

Jamrozik, J., Schaeffer, L.R., Liu, Z. & Jansen, G. 1997. Multiple trait random regression test day model for production traits. *Interbull Bulletin No. 16*, 43-47.

Kirkpatrick, M., Hill, W. G. & Thompson, R. 1994. Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. *Genet. Res. Camb. 64*, 57-69.

Mäntysaari, E.A. 1999. Derivation of multiple trait reduced rank random regression model for the first lactation test day records of milk, protein and fat. *50th Annual Meeting of EAAP*, Zurich, August 22-26, 1999.

Neumaier, A. & Groeneveld, E. 1998. Restricted maximum likelihood estimation of covariances in sparse linear models. *Genet. Sel. Evol. 30*, 3-26.

Reents, R., Jamrozik, J., Schaeffer, L.R. & Dekkers, J.C.M. 1995. Estimation of genetic parameters for test day records of somatic cell score. *J. Dairy Sci. 78*, 2847.

Reents, R., Dopp, L., Schmutz, M. & Reinhardt, F. 1998. Impact of application of a test day model to dairy production traits on genetic evaluations of cows. *Interbull Bulletin 17*, 49-54.

Strabel, T. & Misztal, I. 1999. Genetic parameters for first and second lactation milk yields of Polish Black and White cattle with random regression test-day models. *J. Dairy Sci. 82*, 2805-2810.

Tijani, A., Wiggans, G.R., Van Tassel, C.P., Philpot, J.C. & Gengler, N. 1999. Use of (co)variance functions to describe (co)variances for test day yield. *J. Dairy Sci. 82*, (Jan.). Online.

Visscher, P.M. & Thompson, R. 1992. Univariate and multivariate parameter estimates for milk production traits using an animal model. I. Description and results of REML analyses. *Genet. Sel. Evol. 24*, 415-430.

White, I.M.S., Thompson, R. & Brotherstone, S. 1999. Genetic and environmental smoothing of lactation curves with cubic splines. *J. Dairy Sci. 82*, 632-638.