The effect of selection on genetic parameter estimates

R. van Dyk[1], F.W.C. Neser[1#] and F.H. Kanfer[2]
[1] Department of Animal Science, University of the Free State, P.O. Box 339, Bloemfontein 9300, South Africa;
[2]Department of Statistics, University of Pretoria, Pretoria 0002, South Africa.

Abstract

A simulation study was carried out to investigate the effect of selection on the estimation of genetic parameters for butterfat production in dairy cattle. It was found that selection leads to a substantial overestimation of fixed effects and variance components and an underestimation of heritability estimates. This effect was attributed to the fact that all information on all animals in the analysis should be available to compensate for the effect of selection.

Keywords: Genetics, selection, animal breeding
[#]Author to whom the correspondence should be addressed; e-mail: NeserFW@sci.uovs.ac.za

Introduction

Genetic variance estimates from selected individuals can be quite different from those in the unselected base population. Lynch & Walsh (1998) stated that REML yields unbiased estimates of additive genetic variance in the base population if the base population consists of unrelated, unselected and non-inbred individuals, and phenotypic data are available for all selected and unselected individuals. Selection may cause bias in the estimation of genetic parameters and breeding values. Data available to animal breeders are invariably derived from herds in which artificial selection has been practiced. Consequently, the assumption of random sampling that is invoked for estimation and prediction is no longer valid (Schaeffer *et al.*, 1998). Simulation is the only apparent method for examining properties of estimators of variances and their ratios (Henderson, 1977). The aim of this study was to compare the fixed effect and genetic parameter estimates for both a selection and non-selection scenario utilizing simulated data.

Material and Methods

A simulated dairy herd was used as the basis for this study. All females generated during the first four years were retained, and after this replacement procedures typical of a dairy enterprise were simulated. One hundred AI sires, all from the same genetic level, were simulated. Ten bulls were used as sires during the first year. During the second year five new bulls were added while five bulls from the previous year were retained. This pattern was continued for 20 years in order to establish strong genetic ties between years. Bulls were thus used on a random non-selected basis. Two different scenarios were simulated, *viz.* with or without selection. This process was repeated on an annual basis for the remaining term. For each simulated cow in the herd the following information was recorded: cow number, fixed effect level at first lactation, production measurement at first lactation, genetic component of the measurement, error component of measurement, sire and dam. In the selection scenario, animals were ranked according to the genetic component of their butterfat measurement and the lowest-performing 25% of all animals were replaced by the best-performing progeny. Cows older than six years of age were also replaced by the remaining best-performing progeny. Selection was done on a yearly basis. In the non-selection scenario, the lowest-performing 25% were replaced with animals drawn at random from the progeny, implying that no attention was given to the performance of the progeny. The comparison between these two scenarios allowed analysis of the effects of selection on the estimation of fixed effects and (co)variance components, while the second scenario also permitted cross-evaluation of the model. The model contained a fixed effect, random genetic and random error components. The fixed effect was regarded as a year or herd-year effect. The model used to simulate the sample was:

$$y_{ij} = f_i + a_{ij} + e_{ij}$$

where y_{ij} represents butterfat,
f_i represents the fixed effect at level i,
a_{ij} represents the random genetic component of animal j under fixed effect i,
e_{ij} represents the random error component.

The random genetic component (a_{ij}) and error component (e_{ij}) had normal distributions with means equal to zero and variances of σ_a^2 and σ_e^2 respectively. (($a_{ij} \sim N(0, \sigma_a^2)$, ($e_{ij} \sim N(0, \sigma_a^2)$). The genetic and error components were statistically independent. This corresponds with the method used by Van Vleck (1993) to obtain pseudo-random values from a normal distribution with a mean of zero and variance of one, and is similar to a Monte Carlo simulation. Tuchscherer & Herrendörfer (1998) and Canavesi & Miglior (1999) proposed similar models to evaluate BLUP estimates. A standard desktop computer with a Linux Red Hat (version 6.0) operating system was used for the simulations, which were programmed in Fortran.

The fixed effect component was calculated as:

$$f_i = c + m(i)$$

where

$m_i = 12.5$ and $c_i = 100$ if $1 \leq i \leq 8$
$m_i = -4.5$ and $c_i = 201$ if $9 \leq i \leq 16$
$m_i = 10.0$ and $c_i = 10$ if $17 \leq I \leq 20$.

The genetic and error components were obtained by generating random numbers from a normal distribution with a mean equal to zero and variances of σ_a^2 and σ_e^2 respectively. The genetic ($\sigma_a^2 = 293$) and error variances ($\sigma_e^2 = 534$) of du Toit *et al.* (1998) were used in the simulation. This implies that a heritability estimate of 0.35 was used.

The simulation had to be repeated several times in order to describe the randomness in the system and to determine the long-term expected effect, which was used for comparisons. Because of logistic constraints, only 50 repetitions were simulated per scenario. The data, consisting of animal number, sire, dam, fixed effect level, and production measurement were analyzed using REML procedures (Meyer, 1995). The estimated fixed effect, predicted breeding values and estimated variance components were determined from this. These results were then compared with the true values used in the simulation.

The model that was fitted for the analysis of the data was the same as that used for the simulation. The relationship between animals was included in the covariance structure in both cases. The model used was a special case of the general model:

$$y = X\beta + Za + e$$

Where y is an $n \times 1$ vector of records, X is an $n \times p$ incidence matrix that relates data to the unknown vector of location parameters β. The vector β contains year as fixed effect. The incidence matrix Z relates the unknown random vectors of the direct breeding value a to y. The unknown vector e contains the random residuals due to environmental effects peculiar to individual records.

Results and discussion

The true and average fixed effect levels, as well as the 99% confidence interval for the true fixed effect levels calculated from fifty simulation repetitions are presented in Tables 1 (non-selection scenario) and 2 (selection scenario). The 99% confidence interval for the non-selection scenario includes the true value for years one to 20, and gives an indication of the accuracy with which fixed effect levels were estimated. It is interesting to note that the estimated fixed effects are included in the 99% confidence interval from years one to four. This is ascribed to the fact that no selection was exercised during this period. The fixed effect levels for the remaining period were substantially overestimated in the selection scenario. In the non-selection scenario, the fixed effect levels remained close to the true levels used in the simulation. These results are depicted in Figure 1 where the true fixed effect and the fixed effect in both the selection and non-selection scenarios are presented.

Table 1 True fixed effect levels, average estimated fixed effect levels and 99% confidence intervals for the true fixed effect level (kg) in the non-selection scenario

			99% Confidence Interval	
Fixed effect level	True level	Average	Lower border	Upper border
1	112.5	112.2	111.5	112.9
2	125.0	125.0	124.4	125.7
3	137.5	137.4	136.8	138.0
4	150.0	150.3	149.6	151.0
5	162.5	154.4	153.3	155.4
6	175.0	167.0	166.0	168.0
7	187.5	185.2	183.9	186.5
8	200.0	197.3	195.9	198.6
9	160.5	157.8	156.6	158.9
10	156.0	152.9	151.7	154.2
11	151.5	148.8	147.8	149.9
12	147.0	145.0	143.9	146.1
13	142.5	141.0	139.9	142.1
14	138.0	134.0	135.7	138.2
15	133.5	132.7	131.4	134.0
16	129.0	127.7	126.3	129.1
17	180.0	179.3	178.1	180.6
18	190.0	189.3	188.2	190.3
19	200.0	199.4	198.2	200.6
20	210.0	209.6	208.4	210.8

Table 2 True fixed effect levels, average estimated fixed effect levels and 99% confidence intervals for the true fixed effect level (kg) in the selection scenario

			99% Confidence Interval	
Fixed effect level	True level	Average	Lower border	Upper border
1	112.5	112.6	111.7	113.3
2	125.0	125.3	124.4	126.2
3	137.5	137.3	136.5	138.1
4	150.0	149.9	149.0	150.6
5	162.5	168.0	166.9	169.2
6	175.0	183.7	182.6	184.9
7	187.5	201.5	200.2	202.7
8	200.0	214.2	212.7	215.8
9	160.5	176.0	174.7	177.2
10	156.0	171.8	170.4	173.1
11	151.5	166.0	164.8	167.2
12	147.0	162.9	161.4	164.4
13	142.5	159.3	157.7	160.9
14	138.0	154.8	152.9	156.8
15	133.5	151.8	150.3	153.2
16	129.0	147.2	145.4	148.9
17	180.0	198.2	196.4	200.1
18	190.0	208.5	206.9	210.1
19	200.0	218.7	217.1	220.3
20	210.0	228.7	227.3	230.1

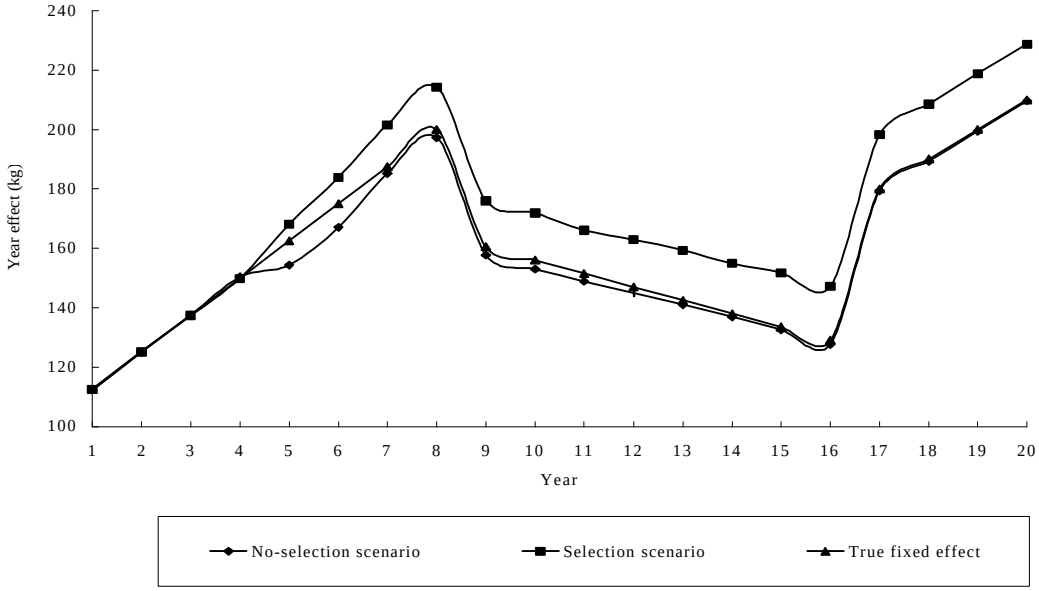

Figure 1 Estimated and true fixed effect levels

There was considerable overestimation of fixed effects under the selection scenario (Table 2 and Figure 1). When selection was excluded from the simulation, the estimate of the fixed effect was much closer to the true fixed effect value. The decrease in years 5-6 under the non-selection scenario is ascribed to chance.

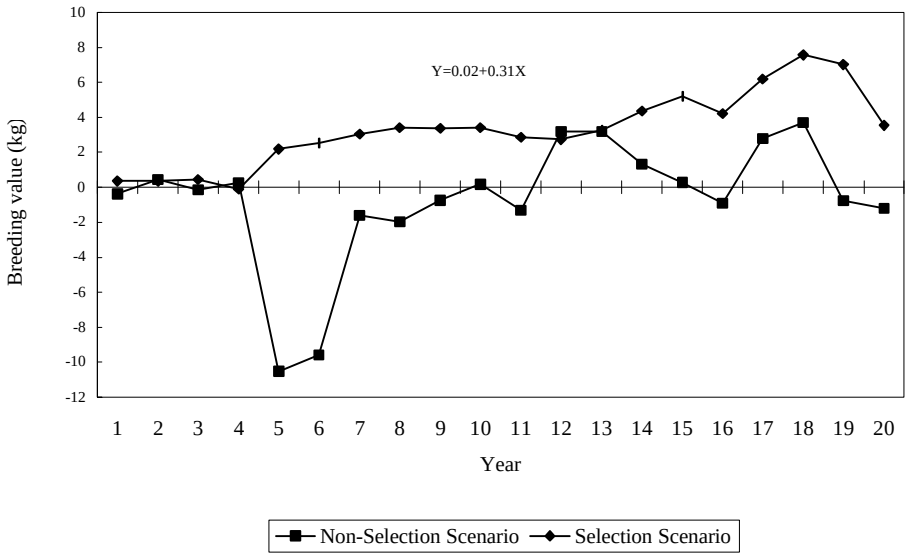

Figure 2 Genetic trends for the selection and non-selection scenario

Figure 2 shows genetic trends for the selection and the non-selection scenarios. The trend under the selection scenario showed a long-term increase of 312 g per year. The long-term trend for the non-selection scenario was influenced by the averages for years 5 and 6, which were excluded from the analysis since it is believed that they can be attributed to chance. No long-term trend was evident, as would be expected under a non-selection scenario.

The estimated genetic and error variances for the 50 repetitions of the two scenarios are presented in Tables 3 and 4. The tables also include the true genetic and error variances used in the simulation as well as the 99% confidence interval.

Table 3 Estimated variance components for the selection scenario

Simulation Round	Genetic Variance	Error Variance	Simulation Round	Genetic Variance	Error Variance
1	95.14	490.64	26	102.90	527.18
2	86.13	517.58	27	117.14	509.07
3	50.12	548.43	28	63.46	541.97
4	80.67	527.43	29	107.51	526.57
5	85.70	521.77	30	85.70	521.77
6	86.07	515.36	31	85.70	521.77
7	74.91	538.06	32	85.70	521.77
8	71.20	552.19	33	119.86	499.06
9	65.97	528.25	34	116.14	500.21
10	72.48	531.88	35	87.25	507.83
11	67.81	533.89	36	103.69	497.18
12	121.84	488.95	37	127.79	507.26
13	105.33	497.60	38	92.09	511.17
14	85.00	508.44	39	64.56	529.08
15	99.29	505.05	40	100.03	521.27
16	59.36	549.51	41	88.86	519.32
17	68.66	503.53	42	55.83	548.71
18	78.62	511.09	43	85.70	521.77
19	54.10	530.88	44	78.70	523.86
20	115.08	504.64	45	55.63	539.76
21	106.39	509.92	46	114.23	495.84
22	83.28	522.35	47	68.75	516.85
23	119.85	493.32	48	91.99	497.90
24	83.12	597.68	49	77.96	528.22
25	37.19	577.09	50	54.37	547.46
			Mean	85.70	521.77
			s.d.	21.43	21.45
			99% confidence interval	76.21-92.74	513.48-529.46

Table 4 Estimated variance components for the non-selection scenario

Simulation Round	Genetic Variance	Error Variance	Simulation Round	Genetic Variance	Error Variance
1	240.89	430.54	26	201.68	447.10
2	197.59	458.28	27	232.17	439.86
3	219.57	447.40	28	199.86	461.33
4	202.51	448.32	29	214.87	448.47
5	206.52	451.07	30	222.12	424.47
6	225.60	459.77	31	193.02	456.71
7	245.35	447.73	32	239.29	437.99
8	258.21	428.68	33	209.50	427.49
9	224.98	448.42	34	239.79	433.58
10	218.44	457.80	35	202.57	440.47
11	227.97	452.67	36	197.63	478.44
12	228.74	443.78	37	220.27	442.41
13	214.56	455.47	38	220.11	447.87
14	213.33	450.42	39	214.58	442.20
15	202.33	447.70	40	188.99	460.51
16	244.10	425.71	41	195.70	463.34
17	218.63	437.06	42	207.63	439.25
18	268.35	395.75	43	225.45	430.90
19	197.16	463.01	44	207.95	430.93
20	208.96	460.99	45	224.74	449.51
21	210.52	444.77	46	238.18	435.82
22	205.52	443.53	47	217.04	430.18
23	234.23	429.12	48	180.23	475.41
24	232.84	433.34	49	224.64	441.25
25	251.74	434.91	50	194.99	421.01
			Mean	219.13	444.32
			s.d.	18.58	14.37
			99% confidence interval	211.70-225.60	438.48-449.86

True genetic variance (293) was substantially underestimated in the selection scenario. The degree of underestimation was less in the non-selection scenario. Although the true error variance (534) was underestimated in both scenarios, the degree of underestimation was less for the selection scenario than for the non-selection scenario. Heritability estimates for the selection and non-selection scenarios are presented in Table 5.

Table 5 Heritability estimates for the selection and non-selection scenarios

Simulation Round	Selection	Non-Selection	Simulation Round	Selection	Non-Selection
1	16.24	35.88	26	16.33	35.60
2	14.27	30.13	27	18.71	31.09
3	8.37	32.92	28	10.48	34.55
4	13.27	31.12	29	16.95	30.23
5	14.11	31.41	30	14.11	32.39
6	14.31	32.92	31	14.11	34.35
7	12.22	35.40	32	14.11	29.71
8	11.42	37.59	33	19.37	35.33
9	11.10	33.41	34	18.84	32.89
10	11.99	32.30	35	14.66	35.61
11	11.27	33.49	36	17.26	31.50
12	19.95	34.01	37	20.12	29.23
13	17.47	32.02	38	15.27	33.24
14	14.32	32.14	39	10.87	32.95
15	16.43	31.13	40	16.10	32.67
16	9.75	36.44	41	14.61	29.10
17	12.00	33.34	42	9.23	29.69
18	13.33	40.41	43	14.11	32.10
19	9.25	29.86	44	13.06	34.35
20	18.57	31.19	45	9.34	32.55
21	17.26	32.13	46	18.72	33.33
22	13.75	31.66	47	11.74	35.34
23	19.55	35.31	48	15.60	33.53
24	12.21	34.95	49	11.75	32.12
25	6.05	36.66	50	12.86	27.49
Mean				14.08	33.01

A heritability of 35% was used in the simulation process. A substantial underestimation of heritability occurred in the selection scenario, as was the case with estimated additive variance. The heritability estimates in the non-selection scenario were more consistent with the true value used in the simulation process.

Product-moment correlation estimates between the true and predicted breeding values are presented in Table 6. The correlations ranged from 0.47 to 0.71 with an average value of 0.61 for the fifty repetitions under the selection scenario, while the correlations ranged from 0.75 to 0.82 with an average value of 0.79 under the non-selection scenario. Van der Werf (1990) also showed that additive genetic variance decreased due to inbreeding, gametic phase disequilibruim (Bulmer, 1971) and covariance among animals after five generations of selection with data used for estimation of genetic parameters that did not include all the relationships and data. Satoh *et al.* (1992) also showed that the correlation increased with an increasing number of generations. Hagger (1991) showed that an increase of up to 55% can be achieved for the correlation between true and predicted breeding values during a selection experiment when more information on the daughters is included.

Table 6 Product-moment correlation estimates between simulated and estimated breeding values derived from DFREML analyses.

Simulation Round	Correlation with selection	Correlation without selection	Simulation Round	Correlation with selection	Correlation without selection
1	0.66	0.82	26	0.64	0.78
2	0.67	0.81	27	0.65	0.80
3	0.55	0.77	28	0.58	0.80
4	0.62	0.80	29	0.66	0.78
5	0.62	0.78	30	0.66	0.78
6	0.70	0.78	31	0.66	0.77
7	0.56	0.81	32	0.60	0.81
8	0.65	0.79	33	0.68	0.81
9	0.59	0.82	34	0.60	0.77
10	0.58	0.78	35	0.59	0.80
11	0.62	0.79	36	0.66	0.78
12	0.71	0.75	37	0.64	0.76
13	0.69	0.76	38	0.60	0.79
14	0.61	0.80	39	0.55	0.78
15	0.57	0.78	40	0.66	0.77
16	0.56	0.78	41	0.69	0.76
17	0.55	0.81	42	0.55	0.78
18	0.51	0.79	43	0.55	0.77
19	0.53	0.82	44	0.65	0.78
20	0.67	0.78	45	0.61	0.80
21	0.64	0.79	46	0.71	0.80
22	0.64	0.78	47	0.59	0.77
23	0.60	0.80	48	0.63	0.78
24	0.64	0.80	49	0.56	0.76
25	0.47	0.79	50	0.51	0.76
Mean				0.61	0.79

In order to account for parental selection, the following should be available: complete pedigrees back to a base population of non-selected, non-related and non-inbred animals (Sorensen & Kennedy, 1984), records on all candidates available for selection (Henderson, 1975; Goffinet, 1983) and knowledge of the selection process and distribution of selection criteria (Henderson, 1975; Im, 1989; Fernando & Gianola, 1990). The first two conditions guarantee that likelihood-based inferences not accounting for selection are the same as those obtained considering selection, regardless of the translation invariance of the selection criterion or its form (linear or non linear) (Gianola & Fernando, 1986; Im, 1989; Fernando & Gianola, 1990). In this study animals were culled before their records could be included in the analysis, as is the case in many dairy enterprises, and the second condition could therefore not be met. Even if the first and second conditions are met, an additional condition of translation invariance of selection criteria must be verified for Henderson's Mixed Model Equations to yield BLUE and BLUP, otherwise unbiasedness does not hold (Schaeffer *et al.*, 1998). The third condition is generally needed when data are missing. Im (1989) showed that if data are missing at random, inferences could be made using the likelihood function without accounting for the missing data process. Otherwise, the missing data process has to be described and included in the likelihood function (Schaeffer *et al.*, 1998).

In agreement with our results, Sorenson & Kennedy (1984) and Van der Werf & de Boer (1990) showed that estimates of genetic variance based on simulated data are unaffected by selection over generations if all data and all genetic relationships since the beginning of selection are included in the analysis. This simulation showed that under certain experimental conditions the REML variance-component estimates are influenced by selection.

Conclusion

A notable feature of the results of this experiment was the differences that existed between the true and estimated values. This appears to be a result of the process of selection the underlying assumptions of the mixed linear model. An assumption of Henderson's Mixed Model Equations is that the expected value of every element of **a** (vector of breeding values) is zero. If the animals under consideration are the result of a long-term selection program, then the expected value of breeding values in later generations could differ from zero. Furthermore, for the records of selected individuals, all variances are reduced and non-zero

covariances are generated between previously uncorrelated effects, such as between **a** and **e** (Schaeffer *et al*, 1998). A model with conditioning on selected base animals cannot be used to obtain unbiased estimates of variance components (as with this study). Conditioning on selected parents requires knowledge about the regression of parents on offspring, and therefore inference has to be made concerning the dispersion of base animals in any case of estimating variances based on their progeny (Van der Werf & Thompson, 1992).

References

Bulmer, M.G., 1971. The effect of selection on genetic variability. Am. Nat. 105, 201-211

Canavesi, F., & Miglior, F., 1999. Effect of De-Regression on Lactation Number on MACE Evaluations: A Simulation study. In: Proc. of the Interbull Meeting Zurich, Switzerland. 22, 49-51

Du Toit, J., van Wyk, J.B. & van der Westhuizen, J., 1998. Genetic parameter estimates in the South African Jersey breed. J. Anim. Sci. 28, 146-152

Fernando, R. L. & Gianola, D., 1990. Statistical inferences in populations undergoing selection or non-random mating. In: Advances in Statistical Methods for Genetic Improvement of Livestock. Eds. Gianola, D. & Hammond, K., Springer-Verlag, New York. pp. 437-453.

Gianola, D. & Fernando, R. L., 1986. Bayesian methods in animal breeding theory. J. Anim. Sci. 63, 217-244

Goffinet, B., 1983. Selection on selected records. Genet. Sel. Evol. 15, 91-98

Hagger, C., 1991. Changes in (co)variance derived properties of animal model breeding values when offspring information became available in a selection experiment. Anim. Breed. Genet. 108, 1-8

Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31, 423-427

Henderson, C.R., 1977. Simulation to examine distribution of estimators of variances and ratios of variances. J. Dairy Sci. 61, 267-273

Im, S., 1989. A note on sire evaluation with uncertain paternity. Biometrics 31, 749-752

Lynch, M. & Walsh, B., 1998. Genetics and analysis of quantitative traits. Sinauer Assoc. Sunderland, Massachusetts, U.S.A.

Meyer. K., 1995. DFREML programs to estimate variance components restricted maximum likelihood using a derivative free algorithm. User notes.

Satoh, M., Nishida, A. & Furukawa, T., 1992. The influence of variation in the generation effect on the accuracy of predicting breeding values – a computer simulation on a closed herd of swine. Anim. Sci. & Tech. 63, 457-461

Schaeffer, L.R., Schenkel, F.S. & Fries, L.A., 1998. Selection bias on animal model evaluation. Proc. 6th World Congr. Genet. Appl. Livest. Prod. 25, 501-508 (Armidale, Australia)

Sorensen, D.A. & Kennedy, B.W., 1984. Estimation of genetic variance from unselected and selected populations. J. Anim. Sci. 59, 1213-1221

Tuchsceherer, A., & Herrendörfer, G., 1998. Evaluation of estimated BLUP in mixed linear models by a designed computer simulation. Proc 6th World Congr. Genet. Appl. Livest. Prod. 25, 585-588 (Armidale, Australia)

Van der Werf, J.H.J., 1990. A note on the use of conditional models to estimate additive genetic variance in selected populations. Proc. 4th World Congr. Genet. Appl. Livest. Prod. 13, 476-479 (Edinburgh, Scotland)

Van der Werf, J.H.J. & de Boer, I.J.M., 1990. Estimation of additive genetic variance when base populations are selected. J. Anim. Sci. 68, 3124-3132

Van der Werf, J.H.J. & Thompson, R., 1992. Variance decomposition in the estimation of genetic variance with selected data. J. Anim. Sci. 70, 2975-2985

Van Vleck, L.D., 1993. Selection index and introduction to mixed model methods. CRC Press Inc. Florida.