

Correcting for Early Within-family Pre-selection in Genetic Evaluation – A Simulation on Rainbow Trout Growth

*M. Janhunen**, *A. Kause**, *H. Vehviläinen**, *H. Koskinen†*, *A. Nousiainen†*

*MTT Agrifood Research Finland, FI-31600 Jokioinen, Finland, †Finnish Game and Fisheries Research Institute, FI-72210 Tervo, Finland

ABSTRACT: Rainbow trout breeding scheme was simulated to estimate how within-family pre-selection of fingerlings and information on culled fish affect genetic evaluation of grow-out weight traits in two environments. Fish for different datasets were randomly sampled (R) or pre-selected and information on culled fish were either individually measured (S+IND), replicated with their family-specific averages (S+AVER), or missing (S-MIS). Variance estimates in R and S+IND did not diverge from simulated values, whereas S+AVER decreased residual variances. Accuracies of EBVs were equally high for R, S+IND and S+AVER. For S-MIS, convergence problems occurred, variance components were distorted, and EBV accuracies were low. Selection bias was consistently expressed by overestimated genetic and common environment variances, and underestimated residual variances. Data adjustment by S+AVER is concluded to sufficiently control for selection bias in genetic evaluation of growth, but for estimation of variances R or S+IND are preferable.

Keywords: culling; fish breeding; selection bias

Thompson (1984); Pollak et al. (1984); Outweltjes et al. (1988)). However, the assumptions of this statistical technique are no longer fulfilled when records of the non-selected i.e. culled animals are missing. Because selection bias may hinder the achievement of intended genetic progress, appropriate means to adjust for pre-selected data are needed.

In this study, a stochastic simulation was used to investigate how within-family pre-selection based on fingerling weight and different precision of information of culled fish affect variance parameters and accuracy of genetic evaluation in different grow-out body weight traits. The specific interest was to compare how the effect of pre-selection on genetic parameter estimates can be accounted for when the observations of fingerling weight for culled fish are either 1) individually measured, 2) replicated as their family-specific averages, or 3) are missing. The design of the simulation was adopted from a Finnish breeding program for rainbow trout, *Oncorhynchus mykiss* (Kause et al. (2005); Janhunen et al. (2013)).

Introduction

Some genetic improvement programs for fish apply a two-stage selection scheme. In this scheme phenotypic pre-selection is first practiced within families based on body size of juvenile fish (usually at id-tagging). At the second stage, brood fish candidates are evaluated and selected for mating based on their estimated breeding values (EBVs) for economically important traits, such as harvest body weight. Pre-selection improves genetic response in the breeding objective, i.e., in traits correlated with the pre-selected trait (Martinez et al. (2006); Sae-Lim et al. (2013)). Other benefits of pre-selection are fixed animal numbers among families, and reduced number of individuals reared up to harvest or reproduction, which lower management costs of the program (Mueller (1984)).

Large family sizes and considerable variation of growth in most cultivated fishes enable high selection intensities to be used within families. However, intense pre-selection may cause a bias in the estimates of (co)variance components and EBVs, unless information on the culled animals is included in genetic analyses (Henderson (1975); Meyer and Thompson (1984)). The impact of selection bias on genetic parameter estimates can be usually taken into account by using a multivariate animal model with the pre-selected trait(s) included (Henderson (1975); Meyer and

Materials and methods

Simulation of data. Half-sib data with 150 full-sib families were generated in a stochastic simulation for a breeding population of rainbow trout using a split-family design. The base population comprised 75 dams and 75 sires, which were assumed to be unrelated and non-inbred and had neither pedigree information nor own phenotypic records. Each parent was mated twice using a partial factorial design. The initial full-sib family size was 100.

Simulated traits were fingerling body weight at id-tagging, at six months of age (BW1), and harvest weight at the freshwater nucleus (BW2) and sea test station (BW2_{sea}) at two years of age. Each trait had a phenotypic mean of 0 and they followed a multivariate normal distribution, in which phenotypic variance was partitioned into three (co)variance components. The simulated genetic, common environment and residual (co)variances were derived from the Finnish breeding population of rainbow trout (Janhunen et al. (2013)). BW1 had a genetic and common environment correlation of 0.46 and 0.37 with BW2 and 0.39 and 0.43 with BW2_{sea}, respectively. The corresponding phenotypic correlations were 0.42 and 0.32. Residual covariance between BW2 and BW2_{sea} was set to zero.

For offspring generation, phenotypic records for each body weight trait were simulated as the sum of an individual's true breeding value (TBV), an early environment effect common to full-sibs (i.e., fish are reared in separate family tanks before tagging), and a residual effect. TBVs were simulated as the average of parents' TBVs plus a Mendelian sampling term.

Data treatments. Four differently treated data (i.e., different pre-selection strategies) were generated from the full simulated data. In random sampling (R), individuals were tagged randomly without selection. Other three data treatments represented within-family threshold selection (S), in which the phenotypic value of BW1 was used as a criterion for pre-selection. That is, only the biggest fish from each family were selected. In each data treatment, the proportion of selected individuals per family was 40% ($i = 0.97$), which was divided so that the 25 biggest fish were retained as brood fish candidates in the freshwater nucleus ($n = 3,750$ fish in total) and 15 fish were chosen to the sea test station ($n = 2,250$ in total).

The pre-selected individuals had their original simulated BW1 and either BW2 or BW2_{sea} records, depending on the rearing environment. In selected data scenarios, the BW1 records of culled fish were either 1) measured individually (S+IND), 2) were replicated as the family-specific averages of the culled fish (S+AVER), or 3) were missing (S-MIS).

Analysis. The estimation of variance components and EBVs for BW1, BW2 and BW2_{sea} was performed using a multitrait animal model. For each trait, the model was: $y_{ij} = \mu + \text{Anim}_i + \text{Fullsib}_j + e_{ij}$, where y_{ij} denotes an observation of an individual i , μ is the fixed overall mean of a trait, Anim_i is the random genetic effect of an individual i ($i = 1 - n$ of fish), Fullsib_j is the random common environment effect of full-sibs modeled without pedigree information ($j = 1$ to 150 family tanks), and e_{ij} is the random residual effect.

Variance components were analyzed using DMU 6.0 software (Madsen and Jensen (2008)), and the EBVs were estimated using MiX99 software package (©Biometrical Genetics, MTT Agrifood Research Finland, www.mtt.fi/BGE/Software). The effect of different data treatments on the accuracy of EBVs was evaluated using Pearson's product-moment correlations between EBVs and TBVs. The correlations were calculated only for the brood fish candidates in the freshwater nucleus.

Each data treatment was simulated 500 times, but the variance component estimates and EBVs were averaged only over the replicates that achieved convergence in the estimation of (co)variance components. The parameter estimates obtained from the alternative data treatments were compared with the simulated parameter values.

Results and Discussion

Adjustment of pre-selected data by average BW1 values of culled fish (S+AVER) could, on average, restore the phenotypic mean in BW1 to zero (consistent with S+IND and R) (Table 1). Instead, the results for the selected data with missing observations (S-MIS) differed from the other data treatments. The number of converged iterations was clearly the lowest in S-MIS (43% of replicates converged). These estimation problems were presumably due to difficulties to disentangle genetic and common environment effects from each other in data with reduced information content. S-MIS also produced substantially higher BW1 means than other data treatments, because no records were available for the small culled individuals.

Table 1. Number of replicated simulations at convergence and average phenotypic means¹ (\pm SD) for tagging weight (BW1) and two harvest weight (BW2 and BW2_{sea}) when alternative data treatments were applied.

Data treatment ²	N of repl.	BW1	BW2	BW2 _{sea}
R	462	0.0 (1.0)	63 (13)	-75 (12)
S+IND	466	0.0 (0.9)	133 (13)	35 (12)
S+AVER	447	0.0 (0.9)	132 (13)	35 (12)
S-MIS	213	12.3 (0.9)	129 (13)	30 (11)

¹ Mean (\pm SD) was calculated across replicates.

² Random sampling (R) and threshold selection within families (S); BW1 records for culled fish are individually measured (+IND), augmented as replicated mean of the culled fish in a family (+AVER), or are missing (-MIS)

There was no observable bias in the average variance parameters of any body weight trait when R or S+IND data treatment was used (Table 2). Instead, S+AVER substantially decreased residual variances from their simulated values, particularly for BW1 and BW2. However, only the h^2 estimate of BW1 showed a notable deviance from the simulated parameter value (+7 percentage points). Similarly to convergence success, within-family pre-selection seriously distorted the estimates of variances in all traits when S-MIS was used (Table 2). For each trait, a considerable reduction was found in residual (and total) variance, whereas concomitant increases occurred in genetic and, in particular, common environment variance. Genetic and common environment variances were upward biased because within-family selection with no records for the culled individuals makes full sibs phenotypically more similar. Because the residual variance of BW1 dropped almost to zero in S-MIS, h^2 and c^2 values became strongly overestimated (+17 and +33 pp, respectively). S-MIS also elevated the h^2 estimate of BW2 from its simulated value (+9 pp).

Table 2. Average estimates¹ (\pm SD) of genetic (V_G), common environment of full-sibs (V_C) and residual (V_R) variances, and heritabilities (h^2) and common full-sib variance ratios (c^2) for body weight traits when alternative data treatments were applied. Simulated parameter values are indicated in bold.

Trait/ Data treatment ²	V_G	V_C	V_R	h^2	c^2
BW1	83	45	128	0.32	0.18
R	83 (26)	45 (11)	128 (13)	0.32 (0.09)	0.17 (0.04)
S+IND	83 (26)	45 (11)	128 (13)	0.32 (0.09)	0.18 (0.04)
S+AVER	83 (26)	46 (11)	85 (13)	0.39 (0.11)	0.21 (0.05)
S-MIS	106 (3)	111 (20)	1 (1)	0.49 (0.05)	0.51 (0.05)
BW2	20524	3951	48818	0.28	0.05
R	20922 (4555)	3861 (1372)	48579 (2804)	0.28 (0.06)	0.05 (0.02)
S+IND	20691 (4357)	3892 (1364)	48812 (2898)	0.28 (0.05)	0.05 (0.02)
S+AVER	20719 (4401)	3928 (1367)	45971 (2785)	0.29 (0.06)	0.06 (0.02)
S-MIS	26269 (4060)	5203 (1692)	38454 (2450)	0.37 (0.05)	0.07 (0.02)
BW2_{sea}	15560	2723	50794	0.23	0.04
R	15161 (4153)	2802 (1413)	51156 (2822)	0.22 (0.06)	0.04 (0.02)
S+IND	15005 (3809)	2825 (1410)	51156 (2822)	0.22 (0.05)	0.04 (0.02)
S+AVER	14965 (3811)	2909 (1390)	50159 (3340)	0.22 (0.05)	0.04 (0.02)
S-MIS	16428 (4230)	4286 (1875)	46465 (2422)	0.24 (0.06)	0.06 (0.03)

¹Calculated across replicates; see N in Table 1.

²For traits and data treatments, see Table 1.

For each body weight trait, the correlations between TBVs and EBVs were equally high for R, S+IND and S+AVER, and the lowest correlations were found in S-MIS (Table 3). The equivalent accuracies in the first three alternatives correspond with the fairly unchanged estimates of genetic variance, and thus suggest the consistency of brood fish candidates selected as parents.

Table 3. Mean Pearson's correlation coefficients¹ between true and estimated breeding values for body weight traits² when alternative data treatments² were applied.

Data treatment	BW1	BW2	BW2 _{sea}
R	0.58	0.66	0.59
S+IND	0.56	0.66	0.59
S+AVER	0.56	0.66	0.59
S-MIS	0.41	0.58	0.54

¹Calculated across replicates; see N in Table 1.

²For traits and data treatments, see Table 1.

In this study, BW1 was only moderately correlated with both harvest weights, which presumably promoted S+AVER to reduce selection bias in each trait. The higher the correlations between BW1 and the harvest traits, the higher the selection bias would be. Further, large family sizes of fish combined with moderate (not too strong) selection intensity within families contribute to the benefit of S+AVER, because the actual means and distributions of the selected trait (here BW1) can then be predicted with a fair degree of precision. In agreement with the present results, however, data augmentation with average records for culled animals was shown to improve the accuracy of genetic

evaluation also in pigs which only have a few offspring per family (Appel et al. (1998)).

Conclusion

The missing data due to within-family pre-selection not only causes difficulties in the estimation process of (co)variance components but also decreases the accuracy of genetic parameters. The biased estimates may lead to incorrect ranking of breeding candidates and thus reduce genetic progress in growth. The replacement of the missing observations with the average of the culled fish in their family was found to be an applicable alternative for the individual measurements of the culled fish. Although data adjustment by replicated averages of culled fish produced somewhat biased variance estimates, the selection bias in genetic evaluation of growth (on-growing weight traits) could be largely accounted for. This method is suitable for large-scale fish breeding programs where within-family selection is performed but practical limits may often prevent individual measurements from all of thousands of culled fish. However, in experimental studies where precise estimates of variance parameters are needed, either random samples from families or individual measurements from all culled fish are preferable.

Literature Cited

- Appel, L. J., Strandberg, E., Danell, B. et al. (1998). *J. Anim. Sci.*, 76:1794–1802.
Henderson, C. R. (1975). *Biometrics*, 31:423–447.
Janhunen, M., Kause, A., Vehviläinen, H. et al. (2013). *J. Anim. Sci.*, 91:5144–5152.

- Kause, A., Ritola, O., Paananen, T. et al. (2005). *Aquaculture*, 247:177–187.
- Madsen, P., and Jensen, J. A. (2008). DMU. Version 6, release 4.7.
- Martinez, V., Kause, A., Mäntysaari, E. A. et al. (2006). *Aquaculture*, 254:195–202.
- Meyer, K., and Thompson, R. (1984). *Z. Tierz. Zuchtungsbiol.*, 101:33–50.
- Mueller, J. (1984). *Genet. Sel. Evol.*, 16:103–120.
- Ouweltjes, W., Schaeffer, L. R., and Kennedy, B. W. (1988). *J. Dairy Sci.*, 71:773–779.
- Pollak, E. J. J., van der Werf, J., and Quaas, R. (1984). *J. Dairy Sci.*, 67:1590–1595.
- Sae-Lim, P., Komen, H., Kause, A. et al. (2013). *Aquaculture*, 372–375:89–96.