### **Interbull Checks of Incoming Data**

*Jette H. Jakobsen and Eva Hjerpe Interbull Centre, Box 7023, SE-750 07 Uppsala, Sweden* 

### Introduction

Interbull conducts international genetic evaluations for production, conformation, udder health, longevity, and calving traits for six breeds and 26 countries around the world. The purpose of international genetic evaluation is for participating countries to be able to compare breeding values of bulls across country borders expressed on own country scales. The evaluation schedule currently counts four routine evaluations where international breeding values are predicted, and two test evaluations where genetic correlations among countries are estimated and international values breeding are predicted. The purpose of a routine evaluation and of a test evaluation is different. The purpose of a routine evaluation is to compute international genetic evaluations that subscribing countries can publish, while the purpose of a test run is to test changes in national and/or international genetic evaluations procedures. To ensure a high quality of international genetic evaluation results a high quality of input data is necessary. A high degree of stability of national breeding values submitted for two consecutive routine runs is therefore required. In case data from a modified national genetic evaluation are submitted for a test run, requirements on consistency with the previous national evaluation are relaxed. However, data submitted for a test evaluation are required to pass genetic trend validation test(s) (Boichard et al., Interbull, 2006). 1995; Graphical representation for participation in International Genetic Evaluations is in Figure 1.



Figure 1. Graphical representation for participation in International Genetic Evaluations; Y=yes, N=no.

### Verification of incoming data

Incoming data are examined by comparing them to previous data. This comparison comprises both population based statistics like standard deviations, correlations and regressions between national breeding values, and statistics related to changes in individual breeding values. These statistics are computed using a verification program. Details of the verification program are described by Klei *et al.* (2002). What is added in this paper is a description of some overall statistics.

Two examples of overall statistics are shown in Table 1 and Table 2, respectively. The example in Table 1 shows almost no change in standard deviation (St. Dev.), and also, regressions and correlations are very close to one. The situation is different for the example in Table 2 where the change in standard deviation is as large as 21.9% between the two consecutive national evaluations, and also, regressions are far away from the expectation of one. This statistic clearly indicates a potential problem in the national data, and the country will be asked to investigate the source of the change.

Table 1. Overall statistics for the evaluations of:: Direct Longevity .

Value based on within year values weighted by Number of Bulls per Year							
	PREVIOUS	CURRENT	CHAI	NGE			
			STATI	STIC			
Mean	102.576	102.554	-0.022	-0.130	standardiz	ed for	st.dev. of the
					mean of P	REVI	OUS
St. Dev.	12.327	12.312	-0.016	-0.001	(relative	to	PREVIOUS
					st.dev.)		
b(PREV CURR) ::	0.995						
b(CURR PREV) ::	0.993						
r(PREV,CURR) ::	0.994						

Table 2. Overall statistics for the evaluations of :: Milk Somatic Cell.

Table 2. Overun Stat	listics for the e	didutions of	WIIK DOIN				
Value based on within year values weighted by Number of Bulls per Year							
	PREVIOUS	CURRENT	CHAI	NGE			
			STATI	STIC			
Mean	0.056	0.033	-0.023	-5.680	standardiz	ed for	st.dev. of the
					mean of P	REVI	OUS
St. Dev.	0.136	0.106	-0.030	-0.219	(relative	to	PREVIOUS
					st.dev.)		
b(PREV CURR) ::	1.261						
b(CURR PREV) ::	0.773						
r(PREV,CURR) ::	0.987						

The overall statistics listed in Table 1 and Table 2 is computed as follows. The means and standard deviation of current and previous national evaluations are computed per birth year of bulls for years with 10 or more observations. In addition, the correlations and regressions of previous on current evaluation, current on previous evaluation and correlation between current and previous evaluation, are computed. The standard deviation is adjusted for the average reliability in order to weigh the standard deviation with the precision. The reliability used for the adjustment is computed as in Formula 1 (i.e. the reliabilities supplied by the national genetic evaluation centre are not used).

The reliability is calculated as:

$$rel_{cal} = \frac{EDC}{EDC + \frac{4 - h^2}{h^2}} \qquad [1]$$

where EDC is Effective Daughter Contribution, and  $h^2$  is heritability.

Note that the regression of current on previous (and previous on current) evaluation are also adjusted for the average reliability. Hence, regression coefficients computed by the verification program can be different from simple linear regression coefficients computed with standard statistical software (e.g., SAS).

The across-year standard deviation (weighted by number of bulls per year) is calculated as shown in formula 2 for the current respective previous national evaluation:

$$stddev_{weighted} = \sqrt{\frac{\sum_{i=1}^{n} \left( \left( stddev_{i} \right)^{2} * nb_{i} \right)}{\sum_{i=1}^{n} nb_{i}}} [2]$$

where  $stddev_i$  is the standard deviation in year *i* adjusted for the average reliability, and  $nb_i$  is number of bulls in year *i*. At least five years with at least 10 bulls per birth year is required for the overall statistic to be calculated. The absolute difference in across-year standard deviation is calculated by subtracting the across-year standard deviation in the previous run from the across-year standard deviation in the current run. Also a relative change in across-year standard deviation is computed by relating the absolute change in across-year standard deviation to the across-year standard deviation from the previous run.

The verification program flags large changes in standard deviations between runs. The relative change can be sensitive to changes in breeding values in small populations. Usually, flagging of a large change is an indication of a possible large change in the REML-estimates of sire standard deviation. A change of more than 5% in REML estimates of sire standard deviation can cause exclusion of data from an Interbull routine evaluation (c.f. Figure 1).

# Check of change in standard deviations between runs

Another program in the Interbull workflow for processing incoming data computes means and standard deviations of national breeding values per birth year of bull if at least two bulls have data in a birth year. The mean and standard deviation is subsequently pooled across birth years. The years that currently are used are 1986-1999 for Holstein and 1981-1999 for other breeds. The pooled standard deviations from the current and the previous run are compared, and a change of between five and 10 % is flagged with one star, while a change of more than 10 % is flagged with three stars. An example of pooled standard deviations for milk, fat

and protein are listed in Table 3. In this example, no changes are above the threshold of 5 % and therefore, no reasons for concerns. A different situation is pictured in Table 4, where change in standard deviation between runs is above 10 %. The change is flagged with three stars indicating a potential problem.

 Table 3. Pooled standard deviation years:
 1986 - 1999.

	Current run	Previous run
POOLED STANDARD DEVIATION: Milk	206.5505	209.1700
POOLED STANDARD DEVIATION: Fat	6.4322	6.5000
POOLED STANDARD DEVIATION: Prot	5.5490	5.6300

Table 4. Pooled standard deviation	years:	1986 -	1999.
------------------------------------	--------	--------	-------

	Current run		Previous run
POOLED STANDARD DEVIATION: SCC	.1292	*** Check previous	.1144
POOLED STANDARD DEVIATION: Mast.	NaN		999.9999

## Check on REML-estimates of sire variances

National evaluations are deregressed within country (Jairath et al., 1998) and **REML-estimates** of sire standard deviations are computed (Sullivan, 1999). The estimated sire standard deviations are compared with the estimates from the previous run. In general, a relative change larger than 5 % is interpreted as a major change in the national evaluation and data from a modified national evaluation can first be included in a test evaluation. An example of the comparison of current and previous sire standard deviations for protein vield are shown in Table 5. All changes in this example were less than 5 % and all data were included for that particular routine run.

**Table 5.** Comparison of sire standard deviation for protein yield.

Country	Current	Previous	Pct.
			Difference
CNT1	18.76033	18.82526	-0.34
CNT2	6.61438	6.76092	-2.17
CNT3	4.80582	4.73110	1.58
CNT4	6.69489	6.69489	0.00
CNT5	6.86996	6.86996	0.00

After passing this step, the checks on national data are finalized and the work flow for international evaluations can proceed with combining data from all countries to estimate genetic correlations for prediction of international breeding values. National evaluation centers that wish to compute REML-estimates of sire standard deviations can download the software from the Interbull ftp-server.

# Check of consistency between data and pedigree files

Data-files sent to the Interbull centre for international genetic evaluations are numbered 010, 015, 016, 017, and 018, for production, conformation (115 for Brown Swiss), udder health, longevity, and calving traits, respectively. File format can be found at www.interbull.org. Pedigree information is submitted together with production records in the 010-files. Data files for functional traits are checked against 010-files for ID-consistency. The main reason for this is to check if all bull IDs in the file for functional traits are present in the file with pedigree information. Bulls with national evaluations for functional traits but not present in the production and pedigree file will not get an international evaluation for the functional traits. Reasons for the IDinconsistency can be that time period for data inclusion differs between traits. One example is that countries typically have saved test-day records for a shorter time than lactation records, so when a test-day model is introduced many bulls appear in the files with functional traits that are not in the production and pedigree file. One possibility to solve this is to include

pedigree information in the 010 file but fill in missing values for number of herds, daughters, EDCs, and national evaluation. In that case, the bulls will get international breeding values for the functional traits but not for the production traits. However, the main reason for this check is to pick up inconsistencies in identifications among files.

### Discussion

The main reason to perform checks on incoming data is to ensure that only high quality input data is included in the international genetic evaluation. A high quality of input data increases the probability of fair international genetic comparisons of dairy bulls. It is the "National responsibility" to participate in a test run before entering the evaluation scheme for the first time and before introducing major changes in the national evaluations.

There is no clear-cut answer to the question when a change is a major change. However, if the change in estimated sire standard deviation (REML) is larger than 5% participation in a test evaluation is usually required.

It is the experience that seemingly small changes in the national evaluation procedures can have large impacts on breeding values. Some guidelines for different types of changes are in Table 6. These examples are not to be treated as definite rules. In case of uncertainty, it is recommended to compare the new set of breeding values with the previous evaluation using the verification program and to compute REML-estimates of the sire standard deviations using the software available on Interbull's ftp-server.

**Table 6.** Examples of changes in national evaluation systems that were required to go through a test evaluation.

Type of Change	Example
Trait Definition	Change from heel dept to foot angle
Models	Change from lactation model to test day model
Genetic Groups	
Genetic Parameters	Re-estimation of genetic parameters
Computer Platform	Change from Unix to Windows
Convergence Criteria	Change of convergence criteria
Data Editing	Change in time period of data inclusion
Pedigree Changes	Change in pedigree of large number of bulls

#### References

- Boichard, D., Bonati, B., Barbat, A. & Mattalia, S. 1995. Three methods to validate the estimation of genetic trend for dairy cattle. *J. Dairy Sci.* 78, 431-437.
- Jairath, L., Dekkers, J.C.M., Schaeffer, L.R., Liu, Z., Burnside, E.B. & Kolstad, B. 1998. Genetic evaluation for herd life in Canada. J. Dairy Sci. 81, 550-562.
- Klei, B., Mark, T., Fikse, F. & Lawlor, T. 2002. A method for verifying genetic evaluation results. *Interbull Bulletin 29*, 178-182.
- Interbull. 2006. Interbull trend validation procedure. http://wwwinterbull.slu.se/service\_documentation /General/Code\_of\_practice/framesidacode.htm. Accessed June 15, 2006.
- Sullivan, P.G. 1999. Appendix: REML estimation of heterogeneous sire (co)variances for MACE. *Interbull Bulletin 22*,146-148.