How number of starts inform about the selection bias when using performers for breeding evaluation: the example of French trotters.

58th Annual Meeting of the EAAP, Session 17, Dublin, Ireland, August 26th-29th 2007.

*Rose A. Blouin C. Langlois B. INRA-SGQA 78 350 Jouy-en-Josas France.
*+ENS-Lyon

Abstract:

For trotters, qualification allows only 40% of horses to participate in races. The question of a selection bias, when using competition data for breeding value estimation, is therefore raised. To analyse this matter we looked at the number of starts of 2 to 5 year-old French trotters born between 1996 and 2000 (58 841). Four variables were studied: -1- a none or all variable (starter/non starter) and three truncations of the number of starts;-2a- all data including zero; -2b- only starters excluding zero; -2c- limited to earning horses. Sire, Sire-mother and animal model were applied using REML. Corrections were made for age, sex, year, breeding area, and category of the breeder according to the number of horses produced. The range of the estimations of heritability are for 1: 0.30-0.56; for 2a: 0.10-0.18; for 2b: 0.02-0.07; for 2c: 0.04-0.12. A rupture clearly appears when taking into account the zero starts or not, indicating two phenomena: being prepared to participate in races and when prepared, to take part in a different number of races. If sources of over estimations can be discarded, the first phenomenon appears highly heritable. The second phenomenon, however, seems highly environmental or trainer dependant. The distribution of the number of starts for horses earning money compared to those not earning anything allows estimating the number of horses prepared to race but that never started. It leads to correct the apparent selection rate from 30% to 86%. This later percentage should not produce a great bias on breeding evaluations based on earnings.

Key words:
Horse # Trotter # selection bias # breeding value estimation

Introduction

Breeding value estimation for trotters allows planning a better selection policy than mass selection on phenotype. However in the two cases the estimation is based on the analysis of public racing results. These are decreasing in number at the different stages of the animal life: qualification, starting in a race, getting earnings, getting a racing time (Langlois and Blouin 2006). In France, one can note that only 40% of the born horses get qualified. The question of a selection bias is therefore highly raised.

In this paper we will first estimate the genetic parameters of different criteria used to describe this selection process.

We will secondly analyse more deeply the question of the number of starts to try to estimate the real selection basis which is probably not the number of born horses but the number of horses prepared to race.

Materials and methods

Data: We considered all French Trotters born between 1996 and 2000 in order to get five promotions older than 5 years in 2006, the year of the study. That was in total 58 841 horses. The racing results were obtained from the "Société d'Encouragement à l'Elevage du Cheval Français" (SECF, http://www.cheval-français.com). Pedigree data were provided by the "Système d'Identification Répertoriant les Equidés (SIRE), the national computer system for Stud-books (http://www.haras-nationaux.fr).

Variables:

1- The starting status. It is an all or none variable which takes the value 0 for non starter and 1 for starter. It will be analysed under a Probit scale.

2- Number of starts. An equivalent deviation is calculated according to the number observed for each class, zero included. The table published by Ollivier (1981) of i=z/p for the Normal distribution is then used (i is the standardised mean of the p percent best animals selected, when z is the ordinate of the standard Normal distribution at the truncation point).

3- Number of starts without the zeros. The same calculations as before are made, but ignoring the zeros.

4- Number of starts of earning horses. The same equivalent deviation is calculated but only for the horses having earnings.

Models:

Sire, sire and dam, and animal models were applied to the data using software ASREML (Gilmour et al., 2002) and VCE (Groeneveld 1996) for some verification.

Fixed effects considered were year (5 levels from 1996 to 2000), sex (2 levels, males and females), breeding area (6 levels, North (62,59,02,80,60); Normandy (76,27,28,61,14,50); Great west with Brittany (29,22,56,35,44,85,17,79,49,53,72); Ile de France-centre (77,78,91,93,94,95,45,41,37,36,86,87,16,23,63,15,43,19); Other parts of France (Great Lyon, South West, South East); Extra metropolitan birth place and breeder's category (4 levels according to the number of horses produced over the period: 1-2; 3-5; 6-8; 9+).

Distributions:

The distribution of the number of starts can be cut in two parts, that of horses earning money and that of horses not earning anything. This latter one can be modeled by $p_n = p_1^n$ where p_n is the probability of zero earning after n starts and p_1 is the probability of being not placed in one race. One can remark tat p_1 can be chosen to get the best fit with the observed values. We choose to fit for the two classes with maximum and minimum number (one) by a log linear function.

Otherwise having the number N_n of placed and N_n' of non placed horses for each number of starts n, it is also possible to simply infer the number N_0 of horses prepared to race but that never started:

$Nn' = \{N_n'/(N_n + N_n')\} \text{ x } N_0$

$N_0 = N_n + N_n'$

The distribution of the number of starts gives therefore the distribution of N_0

It can be estimated trough the mode.

N_0 is then estimated in two different ways

Table1: Estimation of heritabilities

Variables	Model	3 y-old		4 y-old		5 y-old		2-5 y-old	
Starting status	Sire	0,456		0,514		0,451		0,450	
	Sire - Dam	0,442	0,370	0,502	0,390	0,440	0,296	0,559	0,430
Number of Starts, zero included	Sire	0,104		0,133		0,095		0,160	
	Sire - Dam	0,097	0,205	0,124	0,215	0,090	0,144	0,148	0,244
	animal	0,123		0,150		0,099		0,176	
Number of Starts, zero excluded	Sire	0,070		0,034		0,023		0,054	
	Sire - Dam	0,067	0,383	0,034	0,230	0,022	0,252	0,052	0,219
	animal	0,070		0,027		0,022		0,047	
Number of starts of earning horses	Sire	0,112		0,056		0,038		0,060	
	Sire - Dam	0,106	0,512	0,053	0,420	0,037	0,312	0,058	0,291
	animal	0,121		0,057		0,040		0,050	

Estimation error is between 1 and 2 per cent for Animal and sire path and between 3 and6 per cent for the maternal path.

Table 2: Significance of fixed effects according to the different variables and models from 2 to 5 years of age

Variables	Model	Year	Sex	Region of birth	Breeders category
Starting status	Sire	*	**	**	**
	Sire-Dam	*	**	**	**
Number of Starts, zero included	Sire	*	**	**	**
	Sire-Dam	**	**	**	**
	animal	**	**	**	**
Number of Starts, zero excluded	Sire	*	**	**	ns
	Sire-Dam	*	**	**	ns
	animal	*	**	**	ns
Number of starts of earning horses	Sire	**	**	**	ns
	Sire-Dam	**	**	**	ns
	animal	**	**	**	ns

Table 3 Corresponding effects

Effect of sex:

Variables	male	female
Starting status	0	**- 0,102 +/- 0,011**
Number of starts	0	**- 1,452 +/- 0,106**
Déviation Number of starts	0	**- 0,080 +/- 0,007**
Number of starts zéro excluded	0	**- 2,077 +/- 0,201**
Déviation Number of starts zéro excluded	0	**- 0,146 +/- 0,014**
Number of starts of earning horses	0	**- 1,869 +/- 0,208**
Déviation earning horses	0	**- 0,138 +/- 0,015**

Effect of the breeding region:

Variables	**North**	**Normandy**	**Great west & Brittany**	**Ile de France & center**	**Rest of France**	***Extra metropolitan***
Starting status	0	0,039 +/- 0,02	**- 0,100 +/- 0,03**	0,007 +/- 0,03	**- 0,090 +/- 0,03**	***- 0,300 +/- 0,08***
Number of starts	0	**1,141 +/- 0,23**	0,350 +/- 0,25	**0,765 +/- 0,32**	- 0,100 +/- 0,28	***- 1,800 +/- 0,72***
Déviation Number of starts	0	**0,052 +/- 0,01**	- 0,020 +/- 0,02	0,027 +/- 0,02	-0,030 +/- 0,02	***- 0,140 +/- 0,05***
Number of starts zéro excluded	0	**2,288 +/- 0,42**	**3,433 +/- 0,44**	**2,376 +/- 0,56**	**1,827 +/- 0,51**	*- 0,450 +/- 1,62*
Déviation non zéro	0	**0,184 +/- 0,03**	**0,261 +/- 0,03**	**0,198 +/- 0,04**	**0,151 +/- 0,04**	*- 0,070 +/- 0,11*
Number of starts of earning horses	0	**1,843 +/- 0,44**	**2,909 +/- 0,46**	**1,472 +/- 0,58**	**1,273 +/- 0,54**	*0,221 +/- 1,73*
Déviation earning horses	*0*	***0,146 +/- 0,03***	***0,220 +/- 0,03***	***0,126 +/- 0,04***	***0,092 +/- 0,04***	*- 0,020 +/- 0,12*

Effect of breeder's category:

Variables	**1-2**	**3-5**	**6-8**	**9+**
Starting status	0	**0,067 +/- 0,021**	**0,074 +/- 0,023**	**0,156 +/- 0,018**
Number of starts	0	**0,523 +/- 0,190**	**0,607 +/- 0,210**	**1,440 +/- 0,169**
Déviation number of starts	0	**0,035 +/- 0,012**	**0,040 +/- 0,013**	**0,094 +/- 0,011**
Number of starts zero excluded	0	0,530 +/- 0,282	**1,332 +/- 0,345**	**1,019 +/- 0,267**
Déviation non zéro	0	**0,041 +/- 0,020**	**0,100 +/- 0,024**	**0,074 +/- 0,019**
Number of starts of earning horses	0	0,356 +/- 0,286	**0,982 +/- 0,355**	**0,708 +/- 0,275**
Déviation earning horses	0	0,034 +/- 0,021	**0,075 +/- 0,026**	**0,050 +/- 0,020**

Results

Heritabilities and fixed effects estimations are recapitulated table 1, 2 and 3

From table 1 one can remark that heritabilities are much higher when zero are included (starting status $0.30<h^2<0.56$; number of starts $0.10<h^2<0.12$). It decreases when zero are excluded (number of starts $0.02<h^2<0.07$; restricted to earning horses $0.04<h^2<0.12$ when discarding the over estimations due to maternal environment. Not being starter seems therefore highly heritable. However, the number of starts when being a starter appears at the reverse very environmentally dependant.
An increase of the estimation of heritability by the maternal component (h^2_d) can be noted for the number of starts:
When zeros are included (2a) $0.14< h^2_d<0.25$
When zeros are excluded (2b) $0.22<h^2_d<0.38$
Restricted to earning horses (2c) $0.29<h^2_d<0.51$
A differential exploitation of the offspring of the same mare clearly appears. It is probably connected with environmental effects related to the breeder of the mare and therefore confounded with her effect.
From table 2 all fixed effects where found significant except the category of the breeder which play a role on the probability of being a starter but none on the number of starts afterwards.
Table 3 ignore the effect of year (fluctuation not given) but clearly indicate that females have a lesser number of starts than males. The results are given in units of number of starts and in units of standard deviation (underlying Normal score)
Figure 1 gives the distribution of the number of start for horses starting and for zero earning horses from 2 to 5 year of age. The same kind of distribution is found for every age and allows the estimation of N_0 (mode of the distribution for starter, 660 here) and N_0' (exponential model, 610 here), estimating the number of horses prepared to race but which never started.

Figure1

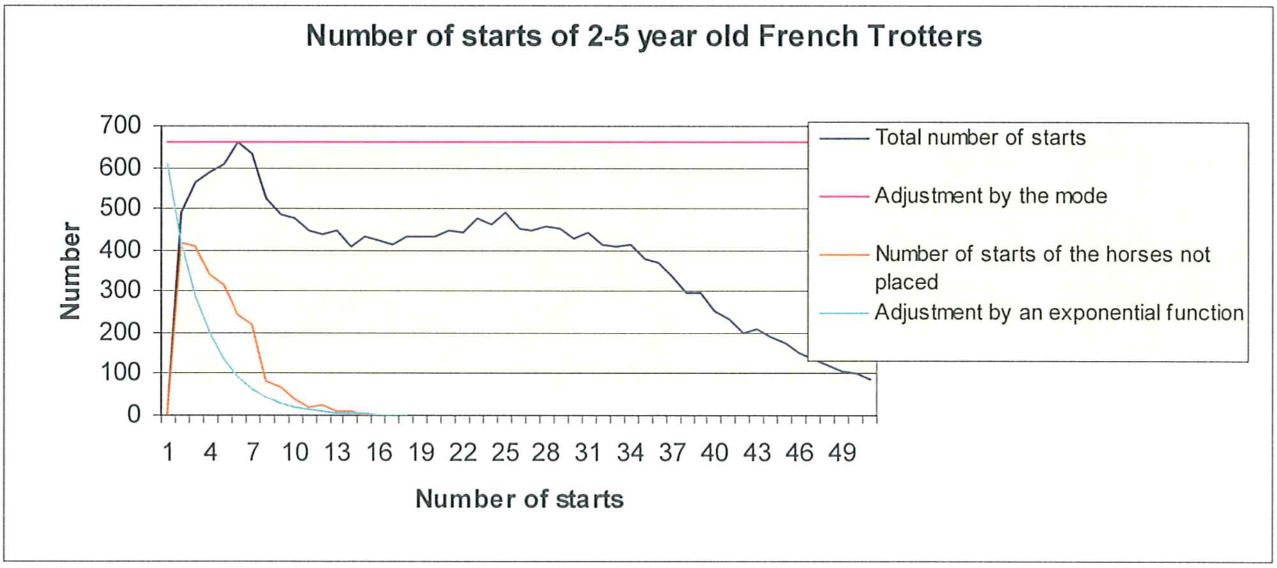


Table 4 summarise the obtained numbers which are converted table 5 for the estimation of different selection rates.

Table 4: Observed and calculated numbers.

N_0: number of prepared horses which never started (mode of Ns)
N_0': id. (Exponential adjustment)
N_s: Number of starters
N_e: Number of horses earning money

Age	N_0	N_0'	N_s	N_e	N_0+N_s	$N_0'+N_s$
2y. old	486	1724	1162	841	1648	2886
3y. old	1636	1313	15062	12685	16698	16375
4y. old	1198	1240	18213	15308	19411	19453
5y. old	1011	1463	14071	11422	15082	15534
2-5y. old	660	610	19781	17573	20441	20390

Table 5: Corresponding selection rates (%)

Age	N_e/born horses	$N_e/(N_0'+N_s)$	$N_e/(N_0+N_s)$	N_e/N_s
2y. old	1.4	29.1	51.0	72.4
3y. old	21.6	77.5	76.0	84.2
4y. old	26.0	78.6	78.9	84.0
5y. old	19.4	73.5	75.7	81.2
2-5y. old	29.9	86.2	86.0	88.8

The amount of horses prepared to race, which did not race is important. According to the two tested hypothesis, it is mostly greater than the half of the number of horses starting but which did not earn anything. However, due to the very great number of earning horses (89% of the starters) this does not impact very much as shown table 5 on the estimated selection rate of the earning horses. When considering the 2 to 5 year old career 30% of the born horses earn some money and they represent in fact a selection rate of 86%. If one would consider the starting horses as the only selection basis this rate would increase to 89%. The rate calculated by Langlois and Blouin (2004) by comparison of BLUP breeding value estimations of earning and non earning horses was 85% for 32% of the born horses. With the present estimation of N_0 or $N_{0'}$, we are therefore in good agreement with previous results.

Discussion

The heritabilities for starting status obtained here by the analysis of variance are very similar as those found by Langlois and Vrijenhoek (2004) by the mother-offspring regression. It is also quite high: $0.30<h^2<0.56$ here and $0.32<h^2<0.43$ and =0.65 for the qualification, estimated before.
The same comparison for number of starts when zeros are excluded gives $0.10<h^2<0.18$ here and was found between 0.07 and 0.11 before. One can remark (table1) much higher estimations by the maternal component (0.14-0.51) indicating some maternal effects on this criteria which could also explain the higher values found by the maternal regression method than by Sire component or the more weighted animal component.
The rupture of the estimations of heritabilities that clearly appears when taking into account the zero starts or not indicates two phenomena:
- One is to be prepared to participate in races or not to be prepared. It seems highly heritable. However, it could be sire dependent but not really heritable leading to a kind of artefact.

To check this problem we conducted the same analyses on the starting status but limiting the data to the stallions having at least n offspring in the file. The results (Figure 2) show a decrease of the estimation of heritability with the increasing number of offspring per stallion.
This fact is in favour of an over estimation of the heritability of this trait due to too much stallions producing too few offspring increasing the chance of being in the same class for a bimodal variable. Additionally, these are very often not really in situation to become race horses. This should be a consequence of the chronic overproduction and of the speculative economy based on the dream to breed a champion without the elementary tools for breaking in and training. However, even with very great number of offspring per stallion the heritability of starting status remains over 0.30. This criteria appears therefore really heritable, this is not an artefact.

The same analyses conducted for the number of starts zeros included did not lead (figure 3) to so much decrease in heritability which is remaining around 0.15. This is an intermediate value between that of the trait starting or not and that of number of starts (zeros excluded) explained by the mixing of the two main phenomena in this criterion.

- The second phenomenon, when prepared to take part in different number of races, seems indeed highly environmental or trainer dependant. Therefore, as recommended before (Langlois and Vrijenhoek, 2004) we confirm that this criteria should be used as correction factor (co variable) for annual or career traits but not as a selection trait. Similar results were found by Arnason (1999) and Svoboda et al. (2005).

Figure 2

Evolution of the estimation of the heritability of the Starting Status between 2 and 5 years according to the minimum number n of offspring per stallion.

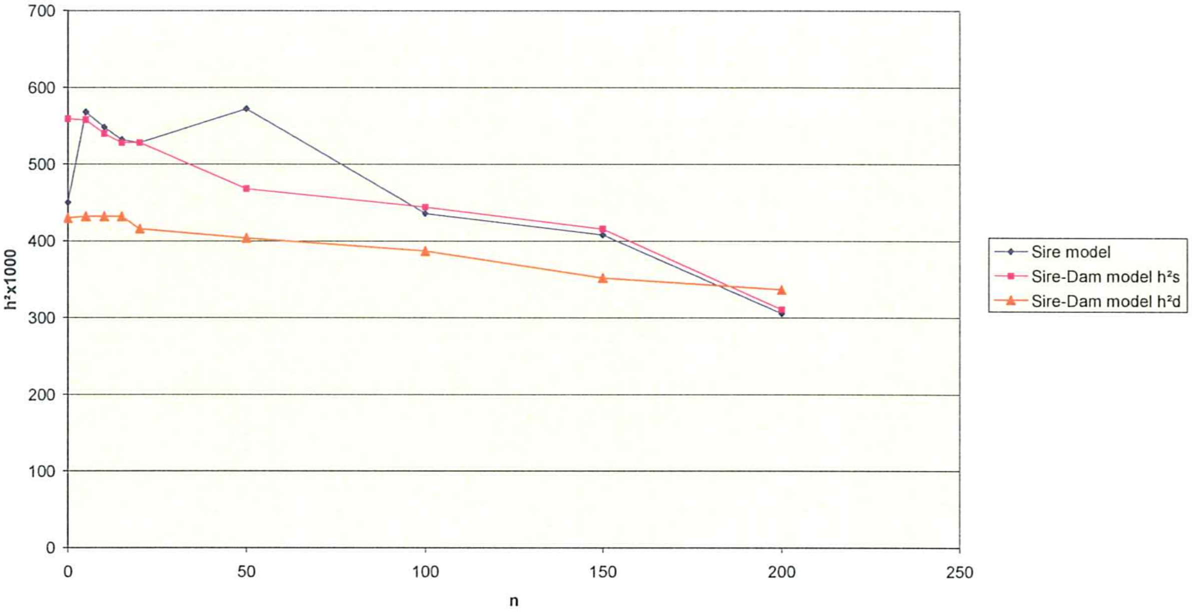
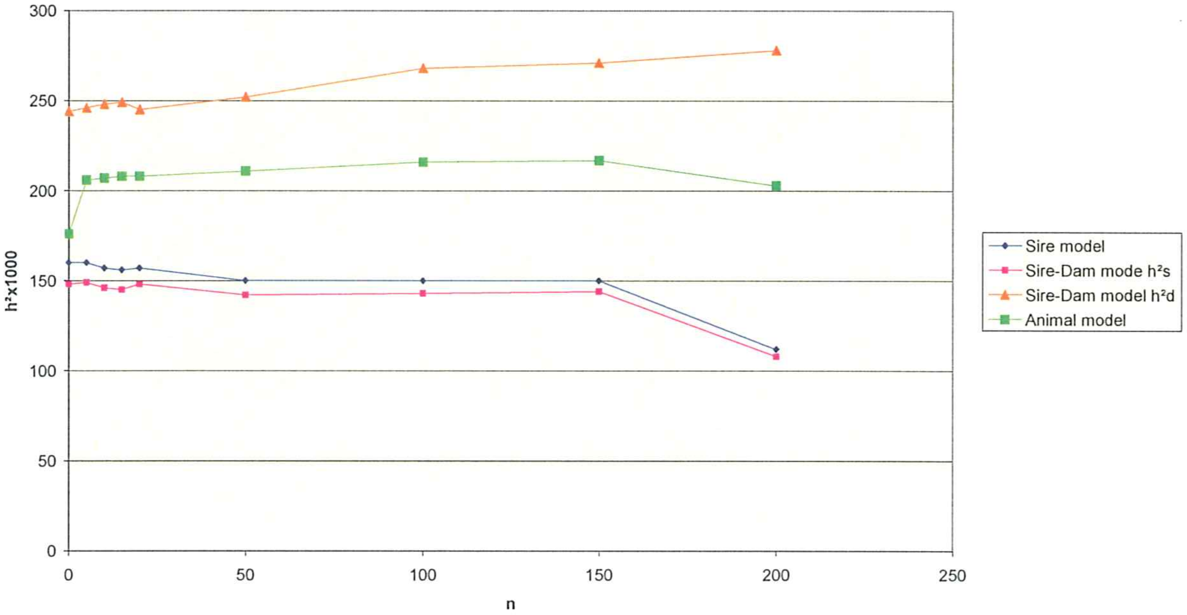


Figure 3

Evolution of the estimation of the heritability of the number of starts from 2 to 5 years of age (zeros included) according to n the minimum number of offspring per stallion



A more thorough discussion of the fixed effects should imply the more detailed results age per age which are not given. We will therefore only discuss the results for region of birth and breeders category for the career traits from 2 to 5 years of age.

The preliminary study by Rose (2006) allowed defining six homogeneous regions regarding starting status and number of starts from the geographical concatenation of birth counties.

Normandy clearly appears as the leading region in terms of starting status where extra metropolitan birth place get the lowest score because French trotters born in foreign countries often do not return in France to compete. North, Ile de France & Centre offer medium chance of being a starter, while great west & Brittany and rest of France lay under the mean. However, when the zero starts are excluded, Great West & Brittany take the lead followed by Ile de France &Centre and Normandy close together and followed by Rest of France and North with the extra metropolitan region remaining at the end. These effects are all significant indicating a different regionalisation of the trait being starter or not compared to the trait number of starts when being starter.

The effect of breeder's category is significant only for starting status and number of starts including the zeros. A clear advantage can be observed with the increasing number of horses produced which is an indicator of a higher professional level of the breeder. This advantage disappear when the zero are excluded, not so big unities (6-8) taking the lead over the very small (1-2) and very big one (9+), even if this tendency was not found significant.

The estimation of the number of horses prepared for racing but which never raced by the mode of the distribution of the number of starts N_0 or by an exponential adjustment of the distribution of the number of starts of non placed horses N_0' are approximate. They suppose that the total potential of starters is fully expressed for maximum of the distribution of starters in the first case (N_0), or in the second case (N_0')that the distribution law proposed to model the distribution of the number starts for none placed horses fit well the reality (this is hardly the case). However, the obtained selection rates (table5) except for the two year-old where the exponential adjustment is difficult to find, are in relatively good agreement. When discarding the disputable 2 year-old results, one can remark that N_0 is much greater for yearly records (table4) than for the career record. This result from the fact that some horses not starting at a given age could start at another age. Seemingly (#600/1200=0.5) that is one half of these none starting horses. The estimated selection rate for earning horses was disputable for 2y.-old but could be estimated between 76 and 77% for 3y.-old, 79 % for 4 -y. old and between 73 and 76 % for 5y.-old leading to an overall selection rate of 86% for the 2-5 y.-old career. Such selection rates will hardly bias breeding value estimations made on earnings.

This is in contradiction with a too much superficial approach evaluating these selection rates from a selection basis constituted by the born horses. We obtain then 22, 26, 19 and 30% respectively.

Conclusion

Because the selection bias when taking in account only placed horses is seemingly not so big as it was previously thought, the recommendations for breeding value estimation made before (Langlois and Vrijenhoek, 2004) can be attenuated.

It was recommended to use two principal traits, the qualification and the career earnings, adding optionally the best time for facilitating the comparison between countries. Number of starts was proposed as important correction factor for earnings (and best time) but not as a selection criterion. Here we confirm these propositions, replacing the qualification by being starter or not which is a heritable criterion. Further studies would lead to estimate the genetic parameters of the variable earning money or not which is probably of the same nature as being a starter, but does not need any additional information as number of starts here and being qualified before, to be treated.

However, the correction of the selection bias expected from the introduction of such none or all variables in the estimation of breeding value of trotters is probably not as important as it was thought.

This are good news given by this study from which it can be concluded that the previous estimations of breeding values made on earnings were not much biased. However, improvement is indeed still possible.

Some further thought is also needed about the meaning of the high heritability of the variable being a starter or not. What is the biological or zootechnical background of this observation? A clear answer to this question is still failing. To enlighten this question the estimation of his genetic correlation with earnings and best time should be properly done. It is important because supposing that being a starter only depends of the horse quality, one could propose to run breeding value estimations only on this criteria. It is indeed available for every born horse, is very heritable and would therefore lead to unbiased estimations with very high accuracy.

But this would be the estimation of the horse quality only in the case of a very high genetic correlation with earnings and best time.

References

Arnason, Th. 1999. Genetic evaluation of Swedish standard-bred trotters for racing performance traits and racing status. Journal of Animal Breeding and Genetics 116, 387-398.

Gilmour A.R., Gogel B.J., Cullis B.R., Wehlam S.J, Thompson R., 2002, ASReml User guide release 1.0 VSN International Ltd, Hemel Hempstead, Hp1 1ES, UK

Groeneveld E., 1996, REML VCE V3.2 User's guide, Institute of animal husbandry and animal ethology, Mariensee 31535 Neustadt, Germany, 52 p.

Langlois B, Vrijenhoek T., 2004 Qualification status and estimation of breeding value in French Trotters Livest. Prod. Sci., 89, 187-194

Langlois B., Blouin C., 2004 Practical efficiency of breeding value estimations based on annual earnings of horses for jumping, trotting, and galloping races in france. Livest. Prod. Sci.87, 99-107

Langlois B., Blouin C., 2007, Annual, career or single race records for breeding value estimation in race horses, Livestock science, 106, 132-141

Ollivier L., 1981, Eléments de génétique quantitative, Actualités scientifiques et agronomiques de l'INRA, éd. Masson.

Rose A., 2006 Estimation des paramètres génétiques du nombre de départs à 2, 3, 4, 5 ans et de 2 à 5 ans chez le trotteur français. Rapport de stage de L3 ENS-Lyon/ INRA-SGQA, 11pp + 3 annexes

Svobodova S., Blouin Ch., Langlois B. 2005. Estimation of genetic parameters of thoroughbred racing performance in the Czech Republic. Anim. Res., 54, 499-509.