Relative Performance Values in Genetic Tests: Alternatives and Their Properties

By R. D. BURDON

New Zealand Forest Research Institute, Private Bag 3020,
Rotorua, New Zealand

(Received 1st April 1997)

Abstract

Relative performance (RP) figures, commonly expressed in relation to a benchmark value of 100 with high values being desirable, are often used as heuristic measures of the performances of provenances or progenies in genetic experiments. A RP figure is defined in terms of the benchmark adopted (e.g. unimproved control mean, experiment mean, or top-performer mean) and the algorithm (the simplest case being a percentage of the benchmark value). Important properties of a RP measure include: the bounds, potential bias, skewness, dependence on convention (e.g. comparing mortality or survival) and biological meaningfulness. The properties of a measure can depend not only on choice of algorithm but also on the properties of the basic data, and the choice of benchmark. Six types of RP measure are considered and their properties are examined in relation to various situations. Measures involving ratios of performance to benchmark levels can have advantages in biological meaningfulness, but if used for incidence of undesirable phenomena may show strong bias and positive skewness. Certain modifications, however, can show better properties. Mortality data in a *Pinus radiata* provenance trial are used as an illustration. Also addressed are issues of applying RP measures to estimates (predictions) of genotypic values rather than just observed entry means, for both single-site and multi-site trials.

Key words: Progeny testing, provenance trials, data analysis, relative performance.

FDC: 165.3; 232.12.

Introduction

In reporting on provenance or progeny tests one should present the results in a form that is easily grasped and interpreted by the reader. To this end, relative performance (RP) statistics can have major heuristic advantages. Moreover, when results are combined from trials at a number of sites, it is desirable that they be expressed for each site on a measurement scale that facilitates the calculation of a fully meaningful measure of combined performance across sites. For example, one may want a scale that gives equal weight to the seedlot differences that are expressed on various sites that differ in growth rate.

An example of using RP in forest tree improvement is found in WRIGHT *et al*. (1966), reporting on a provenance trial replicated across sites. For each site, provenance means were expressed as percentages of the overall mean, at that site, in order to facilitate interpretation of the results. Such RP values can obviously be averaged over sites, and the consistency of RP may well be apparent from casual inspection. The measure of WRIGHT *et al*. was applied to growth data, which would have shown essentially normal distributions and a reasonably narrow range of RP values. As a result, the RP values showed reasonably normal distributions. Very often, however, the data distributions are non-normal, being skewed and/or with variances related to means, as with binomial (0-or-1) data or with visual scores on scales with prescribed bounds. BURDON *et al*. (1992), in the course of reporting results, proposed a solution, based in relation to trial means, to address the problems posed by such data properties.

A different approach to obtaining RP values is embodied in HATCHER *et al.* (1981). This is based on performance relative to the observed range of values, two alternatives mentioned being (1) the extreme range of entry (provenance, progeny or clone) means and (2) the interval of two standard deviations about the experiment mean, which was their preferred option.

Since the paper of HATCHER *et al*. (1981) little, if anything, has apparently been published on methodology for calculating RP values in forest genetic trials, apart from the solution proposed by BURDON and BANNISTER (1985) and modified by BURDON *et al*. (1992).

In this paper, the solution of BURDON *et al*. (1992) will be extended into a more generalised form, with two possible modifications, and the properties of alternative RP measures will be compared. An actual-data example will be used. There will follow a discussion of some issues that arise in inferring genotypic values from entry means and in extending RP measures from single trials to replication over multiple trials.

Towards a Generalised Approach

Addressing one trial at a time, I propose a more generalised specification of a RP statistic in terms of:

1. Benchmark(s)
2. Algorithm.

The combination of the benchmark(s) and the algorithm will effectively set the bounds for the RP values. I will cover the ideal properties of a RP measure, some alternative benchmarks that might be used, alternative algorithms and the properties of the resulting RP values, and some guidelines for choosing of a combination of benchmarks and algorithms.

Benchmarks

Obvious possibilities for choice of benchmark (denoted X*, in practice normally re-set to 100) include:

(i) Trial mean ($\bar{X}$).

(ii) Mean for an *a priori* control, which may represent the genetic material corresponding to a default option for genetic improvement, e.g. an unselected base-population lot in a progeny trial, or a land-race lot in a provenance trial.

(iii) The mean for the top-ranking entry, immediately imposing an upper bound of 100. However, this value, along with any other single-entry mean, will typically have lower precision than a trial mean.

(iv) Conjointly, the top- and bottom-ranking means (set to 100 and 0 respectively) which are embodied in the equation 2 of HATCHER *et al*. (1981). As mentioned earlier, a variation, which

was adopted by HATCHER *et al.*, is a dual benchmark of two standard deviations either side of the trial mean, which then sets the range.

Ideal Properties of Algorithm

The properties of an algorithm are essentially those of the resulting RP statistics. On purely heuristic grounds it is highly desirable for RP values to be positively associated with desirability. In addition to its heuristic value the ideal properties of any RP statistic would be:

1. *Independence of convention*. The potential significance of convention is illustrated by comparing, say, 94% survival vs 97% survival, which at once represents a small percentage difference in survival yet a difference by a factor of two in mortality.

2. *Biological meaningfulness*. A wide difference in RP scores should reflect a difference of large practical importance. The issue, however, is not always straightforward. The survival difference cited above may have little practical significance itself. On the other hand, if the data set is adequate the difference may provide a clue to important survival differences that could arise if establishment conditions were less favourable on the site(s) concerned. The issue here is one of the magnitude of the resolution of entry differences, which may be reflected in proportional variation, rather than the precision of the resolution, although the magnitude and precision must be interrelated to some degree.

3. *Freedom from bias*. Ideally, the mean RP value for the trial ($\overline{RP}$) should correspond to the RP of an entry corresponding to the trial mean (RP mean or RP $\bar{X}$).

4. *Normal distribution* (if the entry means show a fundamentally normal distribution themselves).

Alternative Algorithms

Considering results for one trial (i.e. site) at a time, we may consider the following algorithms:

(i) Simple percentage ratio of entry mean to raw benchmark value (X*). In the straightforward case

$$RP = 100\ X/X^* \qquad (1)$$

where X = mean for entry in question

so that RP = 100 X/($\bar{X}$) if the overall mean for the trial is adopted as the benchmark.

This is most satisfactory where there are no implied upper or lower bounds to the raw values and where the highest raw values are the most desirable.

Where the phenomenon is inherently undesirable, as with the incidence of mortality, disease or stem malformation, an inverse function becomes appropriate:

$$RP = 100\ X^*/X \qquad (2)$$

(ii) For a variable with or without definite bounds, the algorithm of HATCHER *et al*. (1981), in which the entry means are linearly re-coded on the 0 to 100 scale bounded by the upper and lower benchmark values. Specifically, we have

$$RP = \frac{X - X0}{X1 - X0} \times 100 \qquad (3)$$

where X0 and X1 are the lower and upper extremes of the accepted range which may be either the extreme X values or arbitrarily (as used by HATCHER *et al*., 1981) the values two standard deviations of entry means above and below the trial mean.

(iii) With definite upper and lower bounds to the raw variable

$$RP = 100 \times \sqrt{\frac{X - Xworst}{X^* - Xworst} \cdot \frac{Xbest - X^*}{Xbest - X}} \qquad (4)$$

where X = mean for entry in question

Xbest and Xworst are the ideal and the worst values corresponding to the respective ends of the measurement/scoring scale (e.g. 9 and 1 or 6 and 1 for stem straightness scores, or 1 or 0 for survival rate).

This is a generalised version of the formula used by BURDON *et al.* (1992) who used the experiment mean as X*.

It may be noted in connection with equation 4:

(a) This formula is self-correcting for sign and magnitude according to whether, by convention, Xbest > Xworst or *vice versa*,

(b) If $X^* \rightarrow$ Xbest, RP $\rightarrow 0$ in equation 4,

(c) If $X^* \rightarrow$ Xworst the values are not meaningful because almost all RP $\rightarrow \infty$,

(d) If X $\rightarrow$ Xbest, RP $\rightarrow \infty$.

(iv) For $X^* \rightarrow$ Xbest may be appropriate to use:

$$RP = 100\ (X - Xworst)/(Xbest - Xworst) \qquad (5)$$

This is just a linear re-coding of the scale to bounds of 0 and 100, but such bounds do not actually correspond to those of HATCHER *et al*. (1981).

(v) If equation 3 yields highly skewed RP values (i.e., with no clear upper bound) there will be strong positive bias ($\overline{RP} >> 100$). To impose symmetrical bounds and remove most of the bias the following modification can be used:

Rewriting the right-hand side of equation 4 as 100 Q, then for Q > 1

$$RP = 100\ [1 + (1 - Q^{-1})] \qquad (6)$$

For Q ≤ 1 RP can still be calculated as for equation 4.

(An alternative, to inverting Q in equation 4 and reversing the sign, would set a lower bound at $-\infty$, which may have some intuitive attraction, but would just invert the skewness and bias and leave an upper bound of zero. Arbitrarily adding 100 to the values would restore a satisfactory upper bound, but seems contrived).

Properties of the Algorithms

The properties of the alternative algorithms, in relation to the ideal, are summarised in *table 1*. The following points may be noted:

1. Where inverse ratios are involved, as for equation 2 and for many situations with equation 4, there tends to be bias ($\overline{RP} > RP\,\bar{X}$), which is associated with positive skewness for RP values which may have no clear upper bound. The positive skewness may be much more marked than the bias.

2. While equation 3 may not actually generate skewness, the RP values may still reflect skewness resulting from scores clustering towards either bound of a binomial or scoring scale. While the use of data transformations may avert such skewness, it can create other problems (see later).

3. Using the two standard deviations about the mean for equation 3 has the obvious advantage of preventing outlier values for entry means from assuming undue weight. However, such outliers are unlikely to be either spurious or irrelevant for the tree breeder. Moreover, it assumes an underlying normality which, while it may often hold for progeny tests, may not exist in provenance trials.

4. The use of equation 6, which may well be favoured in order to assure symmetrical bounds if the trial mean is the bench-

Table 1. – Summary of properties of the alternative Relative Performance statistics considered.

Algorithm (Eq. no.)	Bounds	Convention-dependence	Bias*	Skewness*	Biological significance's
1	0, undefined	Yes	+	0	Usually good
2	0, ∞	Yes	++	++	May be enhanced
3	0, 100	No	0	0	Potentially spurious
4	0, ∞†	No	++	+	Intermediate between Eqs. 1 & 2
5	0, 100	Yes	+	0	Generally good where indicated
6	0, 200†	No	Much reduced	Much reduced	Potentially improved over Eq. 4

Pluses represent both sign and potential magnitude of sub-optimal feature.
*) Not expected for any algorithm if the benchmark is the top-performer mean or better.
†) Top value set to 100 if the benchmark is the top-performer mean.

mark, is inherently artificial, and does not completely eliminate bias.

5. Biological significance of RP values, while inherently good for equations 1, 2, 4 and 6 (according to the case), may not be closely related to statistical resolution of entry differences. Large main effects of replicates tend to damp down the range of RP values, even if lack of replicate x entry 'interaction' still leaves very efficient resolution of entry differences. Nevertheless, if replicate effects are so important, a modest range of RP values seems realistic.

6. Equation 3 imposes a fixed range of RP values regardless of the biological significance of the differences.

7. The biological significance of large RP differences generated by inverse functions (equation 2 and, to a lesser extent, equations 4 and 6) may be arguable, and such strong resolution of RP differences may come at the cost of considerable bias and skewness.

8. No RP algorithm can overcome fundamental limitations of the data. For instance, with very low mortality and few individuals per entry, small random variations in numbers of dead trees can generate large but essentially spurious variations in RP values for mortality.

Actual-data Example

Various of the points made earlier are illustrated (*Table 2*) by mortality/survival data in a *Pinus radiata* provenance trial (BURDON *et al.*, unpubl.). Sixteen provenance seedlots were each represented as 72 trees planted. There was no *a priori* expectation that seedlot means should be normally distributed, and for present purposes block-replicate effects are deemed to be immaterial. In most respects the data properties make this a fairly extreme example, which serves to illustrate the limitations as well as the advantages of the various RP measures.

With the high overall survival RP values based on survival ratios (Equation 1) showed a relatively narrow range, but they were unbiased and introduced no additional skewness. (In this case Equation 5 gives the same results as Equation 1). By contrast, the mortality ratios (Equation 2) gave wide variation in RP values but strong bias ($\overline{RP} >> RP\ \overline{X}$) with strong associated positive skewness. Using equation 4, which considers both mortality and survival conjointly, has given intermediate results in terms of the spread of values and the degree of bias and skewness. Using equation 6, which modifies equation 4 so as to impose symmetrical bounds about the experiment-mean benchmark that was used, has reduced but not eliminated the bias, and has tended to reverse the skewness. Using some alternative benchmarks for equations 4 and 6 (details not shown), while strongly affecting ranges and means for RP values, did not materially affect bias and skewness.

The lack of bias using equation 3, which was simply a linear re-coding, is illustrated.

The RP values based on ratios of mortality rates (Equation 2 especially) reflect strong impacts of small random variations as mortality approaches nil. This exemplifies how limitations of the basic data cannot be overcome by use of RP measures.

Results for height growth (details not shown) gave RP values with very satisfactory properties using just equation 1.

Estimation or Prediction of Genotypic Values

There is always the issue of deriving from phenotypic values (in this case entry means) the best estimate or prediction of genotypic values. Where trials are replicated among sites this task often becomes more complex.

Single-site Case

For a single site, with numerous entries, multiplying the entry-mean departures from the overall site mean by the heritability (i.e., repeatability) of the means will give the best available estimate of each entry's genotypic value. This gives for each entry a 'shrunken estimate' (cf EFRON and MORRIS,

Table 2. – Seedlot means and relative performance values for survival (mortality + survival = 100%) in a *Pinus radiata* provenance trial. Note: equation 5 is equivalent in this situation to survival % (third column).

Lot No.	X (%)		Relative Performance (Equation No.)				
	Mortality	Survival	1†	2‡	3	4	6
1	4.2	95.8	112	348	94	197	149
2	1.4	98.6	115	1044	100	347	171
3	1.4	98.6	115	1044	100	347	171
4	4.2	95.8	112	348	94	197	149
5	19.4	80.6	94	75	61	84	84
6	16.7	83.3	97	87	67	92	92
7	30.6	69.4	81	47	36	62	64
8	8.3	91.7	107	174	85	137	127
9	9.7	90.3	106	149	82	125	120
10	13.9	86.1	101	104	73	103	103
11	20.9	79.1	93	69	57	80	80
12	27.8	72.2	84	52	42	66	66
13	47.2	52.8	62	31	0	44	44
14	5.6	94.4	110	261	91	170	141
15	8.3	91.7	107	174	85	137	127
16	12.5	87.5	102	116	76	109	109
$\overline{X}$	14.5	85.5	•	•	•	•	•
X*	•	•	85.5	14.5	•§	85.5	85.5
RP $\overline{X}$	•	•	100	258	71	100	100
$\overline{RP}$	•	•	100	258	71	144	112

†) X = Survival
‡) X = Mortality
§) Dual benchmarks of 47.2% and 98.6%

1977). For very finite numbers of entries, or for major imbalance in the classification, such shrunken estimates require more complicated algorithms, but their characteristics are similar. If, for heuristic purposes, one wants to calculate RP values, then it may be appropriate to calculate them on the basis of shrunken estimates, especially when the classification is markedly unbalanced. Applying this approach, it may be noted, will change X1 and X0 used in equation 3, but not Xbest and Xworst in equations 4 to 6.

Extension to Multi-site Trials

If a trial is replicated across several sites, the situation often becomes more complicated. Deriving and combined across-sites RP values can be complicated by several factors, notably various forms of genotype-site interaction (including differential expression of genetic variation), the imperfect precision of entry means within sites, among-site differences in precision of entry means, and any additional imbalance in classification. If there is no genotype-site interaction, equal replication at all sites, and fully homogeneous variances among sites, RP values for average performance across sites will be very satisfactory. Important departures from these ideal conditions however, can make RP values at particular sites of interest in their own right, and raise questions of whether and how RP values as such can be combined satisfactorily over a set or sub-sets of sites.

There are numerous statistical techniques for investigating the various aspects of genotype-environment interaction (e.g. KANG and GAUCH, 1996), including characterising the respective roles of genotypes and environments in generating the interaction. I will not attempt any general coverage, but will make a few comments on the value of site-by-site RP values and the problems of combining them across sites. Looking at RP values for individual entries at a number of sites, it may become readily apparent which entries are behaving consistently or inconsistently. Proper statistical tests, however, will certainly be needed, and several approaches, including that of FINLAY and WILKINSON (1963) (see also WEBER *et al.*, 1996) have considerable heuristic value of their own.

For combining RP values across sites among-site variation in precision of entry means can be a problem. Averaging values based on shrunken means may appear to accommodate this problem, but it does not in itself use the full corroborative value of performances on the various sites. In the extreme situation where between-site genetic correlations for performance are zero no such sacrifice of information would be involved, but in that case each site (or site category) should in principle become a self-contained breeding zone.

The use of a 'site-trait' selection index, which treats the expression of a trait at each site as a separate trait (BURDON, 1979), and genotype-site interaction for a trait as departures from between-site genetic correlations of +1, will use the corroborative evidence afforded by results from various sites. Unlike the more conventional approach to addressing genotype-environment interaction, with genotypes and sites as the two main effects in factorial combination, the use of the site-trait index readily accommodates certain complex heterogeneities among sites in variance and covariance structures. It also provides a powerful framework for predicting gains under alternative patterns of regionalisation. The expression of genotype-site interaction in terms of departure from between-site genetic correlation of +1 is also embodied in the approach of WHITE and HODGE (1989), which uses BLP (Best Linear Prediction) for purely random effects, or otherwise BLUP (Best Linear Unbiased Prediction, used where fixed effects are

involved), to achieve shrunken estimates of breeding values with severely unbalanced classifications.

Where estimates of between-site genetic correlations are not particularly meaningful, as in the case of a very finite number of provenance lots in which the means may not even be normally distributed, attempting to use the corroborative evidence that is implied by between-site covariances/correlations may be questionable. Weighting RP values by heritabilities or repeatabilities of entry means ($h^2_{\bar{g}}$) at the respective sites is an option if a measure of the overall resolution of entry differences is wanted. In this connection, very low or even negative estimates of repeatability of means present a conundrum. In themselves, these estimates imply that the entry means are essentially worthless. Yet experience (e.g. CARSON, 1991) shows that apparent zero repeatabilities of means for a limited number of entries (reflected in F ratio ≈ 1 or < 1) are frequently associated with rankings that concur well with those observed elsewhere. It appears that such cases often reflect random overdispersions of within-entry errors, which can create a paradox whereby the poorer the estimated precision of the means the lower the true errors may be. A suggested approach is set an arbitrary lower bound, say 0.2 or 0.3, to the $h^2_{\bar{g}}$ weightings used for values at individual sites.

Under the approach of HATCHER *et al.* (1981) (Equation 3), RP values averaged over sites will very often be all inside the bounds 0 and 100. Given estimation errors within experiments, an entry that is consistently best may not always top the rankings. Thus while such an entry may rank top overall, its RP scores may average less than 100 across sites. Likewise, a consistently worst entry may average greater than zero. This can be addressed by re-scaling the across-site entry means for RPs to a range of 0 to 100, although that does not in itself accommodate variations among sites in the resolution of entry differences. It must be realised, however, that this general approach, while it may often be a good and heuristically convenient approximation, is only an approximation relative to BLP (of which the selection index is a special case for essentially balanced classifications) or BLUP. It is clearly possible to combine RP values obtained from equations 4 and 6 in the same way, (re-scaling across-site average RP values to a range of 0 to 100) which may have significant advantages for coping with certain data properties.

With multiple sites unbalanced or incomplete representations of entries across sites may impose constraints on choice of benchmarks. If particular entry means are used as benchmarks any such entry will need to be well represented throughout.

The case for basing RP values on shrunken estimates of genotypic values applies to across-sites estimates just as it does to within-site estimates.

Use of Data Transformations

Where data properties indicate the use of a transformation (e.g. square root or log) there is another potential complication. While using a transformation may give more satisfactory analysis of variance it may complicate prediction of genetic gain, and while reverse-transformed averages may inherently be biologically meaningful, they are liable to contain some bias (e.g. JANSSON and DANELL, 1993), which will tend to be inconsistent in the case of shrunken estimates in a very unbalanced classification. The choice of whether to base RP values on means of transformed values, or reverse transformations thereof, may not always be straightforward.

Conclusions

1. Choosing an appropriate algorithm can help accommodate some awkward properties in the basic data in order to give RP values with satisfactory statistical properties.

2. Use of ratios rather than linear re-coding may have major advantages in biological meaningfulness.

3. For incidence of undesirable phenomena, however, the desirability of using ratios may be offset by bias and introduced skewness of RP values.

4. Algorithms using ratios can nevertheless be readily modified to alleviate the bias but not actually eliminate it.

5. Choice of benchmark may restrict choice of algorithm and will affect the range of RP values, but will generally not affect bias or skewness.

6. For single sites RP values may be readily computed for predicted genotypic values rather than just observed mean performances.

7. For multiple sites RP values may help to convey entry differences in consistency of performance, but choice of benchmark may be more constrained.

8. Deriving combined across-site RP values may often be feasible, but is liable to be complicated by less-than-ideal data properties.

Acknowledgments

I thank Dr. C. J. A. SHELBOURNE, Dr. C. T. SORENSSON, and YE GUOYOU for helpful comments on the drafts and Prof. D. LINDGREN for helpful comments on the submitted manuscript.

References

BURDON, R. D.: Generalisation of multi-trait selection indices using information from several sites. NZ J. For. Sci. **9**: 145–152 (1979). — BURDON, R. D. and BANNISTER, M. H.: Growth and morphology of seedlings and juvenile cuttings in six populations of *Pinus radiata*. NZ J. For. Sci. **25**: 123–134 (1985). — BURDON, R. D., BANNISTER, M. H. and LOW, C. B.: Genetic survey of *Pinus radiata*. 2: Population comparisons for growth rate, disease resistance, and morphology. NZ J. For. Sci. **22**: 119–137 (1992). — CARSON, S. D.: Genotype-environment interaction and optimal number of progeny test sites for improving *Pinus radiata* in New Zealand. NZ J. For. Sci. **21**: 32–49 (1991). — EFRON, B. and MORRIS, C.: STEIN's paradox in statistics. Scientific American **236**(5): 119–127 (1977). — FINLAY, K. W. and WILKINSON, G. N.: The analysis of adaptation in a plant-breeding programme. Austr. J. Agric. Res. **14**: 742–754 (1963). — HATCHER, A. V., BRIDGWATER, F. E. and WEIR, R. J.: Performance level – Standardised score for progeny test performance. Silvae Genet. **30**: 184–187 (1981). — JANSSON, G. and DANELL, O.: Needs and benefits of empirical power transformations for production and quality traits in forest tree breeding. Theor. Appl. Genet. **87**: 487–497 (1993). — KANG, M. S. and GAUCH, H. G. (Eds.): Genotype-by-Environment Interaction. CRC press, Boca Raton, FL, USA (1996). — WEBER, W. E., WRICKE, G. and WESTERMANN, T.: Selection of genotypes and prediction of performance by analysing genotype-by-environment interaction. Pp. 353–372. In: KANG and GAUCH (1996). — WHITE, T. L. and HODGE, G. R.: Predicting Breeding Values with Applications in Forest Trees. Kluwer (1989). — WRIGHT, J. W., PAULEY, S. S., POLK, R. B., JOKELA, J. J. and READ, R. A.: Performance of Scotch pine varieties in the North Central Region. Silvae Genet. **15**: 101–110 (1966).