Theor Appl Genet (1993) 87:60–64

THEORETICAL AND APPLIED GENETICS

Comparing test cultivars using reliability functions of test-check differences from on-farm trials

K. M. Eskridge [1], **O. S. Smith** [2], **P. F. Byrne** [3]

[1] Department of Biometry, University of Nebraska, Lincoln, NE 68583-0712, USA
[2] Pioneer Hi-bred International, Johnston, Iowa, USA
[3] Department of Agronomy, University of Missouri, Columbia, MO, USA

Received: 3 December 1992 / Accepted: 1 March 1993

Abstract. An approach to selection is proposed that is based on the probabilities that a test cultivar outperforms a check by more than an amount d in a future environment. The function that gives these probabilities for all possible values of d is called the reliability function. When d = 0, the value of the reliability function is the probability that the test cultivar outperforms the check. The method is illustrated using data from on-farm maize (*Zea mays* L.) strip test trials grown cooperatively by Pioneer Hi-bred International and farmers. Results indicate that reliability functions are useful for evaluating how test cultivars perform relative to a check across a range of environments since the location, slope and shape of the reliability function may be used to describe a test cultivar's performance, similarity to the check and stability, and identify environments where the test cultivar has performance problems.

Key words: Selection – Stability – Reliability – On-farm trials – Strip testing

Introduction

An important consideration in applied plant breeding is the development of new germ plasm that consistently outperforms current cultivars in environments where they are normally grown. As such, many plant breeding programs are placing increasing emphasis on predicting cultivar performance in a broad range of customers' environments and using selection methods that emulate farmers' cultivar selection processes. Thus, trends have been toward (1) more sites with simpler designs at individual sites causing less balance, and (2) the use of easily understood methods that integrate data across many different tests and provide an understanding of the environmental and management factors that contribute to a cultivar's instability (Bradley et al. 1988).

In response to these trends in breeding programs, pairwise mean comparison methods have been proposed as a means to compare cultivars with checks (Jones 1988; Bradley et al. 1988). These approaches compare cultivars with checks using methods that make statements regarding 'true' mean trait values over a population of environments. Such pairwise mean comparisons maximize the number of locations in the comparisons, avoid the problem of imbalance that occurs when differing sets of cultivars are included in different tests and are useful for making comparisons in the presence of genotype-environment (GE) interaction. However, pairwise mean comparisons may not provide information most relevant to decision makers. A major concern of most decision makers, be they farmers, plant breeders or management, is the identification of test cultivars that have a high probability of outperforming the check. Pairwise mean comparison methods may provide little insight regarding the probability of a test cultivar outperforming the check, and making selections based on this probability has been suggested (Eskridge and Mumm 1992). This probability has many of the advantages of pairwise mean comparisons, is easy to understand, and emulates farmers' perceptions of what is important. However, the use of the probability of outperforming a check is limited because (1) it is not useful in comparing the responses of different test cultivars with respect to their similarity to the check over environments, (2) it does not aid the plant breeder in

Contribution from the Agric Res Div, Univ Nebraska, Lincoln, Neb. Paper No. 9913.
Communicated by G. Wenzel
Correspondence to: K. M. Eskridge

identifying environmental factors that may cause a test cultivar's performance to fall below the check and (3) two test cultivars may have nearly identical probabilities of outperforming the check, but one may be clearly superior.

One approach that circumvents these limitations of the probability of outperforming the check is by estimating and graphing the reliability function of each test cultivar (Nelson 1982). This function gives the probabilities that a test cultivar outperforms the check by more than an amount d for all possible values of d. The reliability function for the ith test cultivar is defined as:

$$R_i(d) = P(Y_i - Y_c > d)$$

where P(.) represents probability, Y_i and Y_c are the responses of the test cultivar and check, respectively, and d is any difference. When $d=0$, $R_i(d)$ is the probability (termed reliability) that the test cultivar outperforms the check.

Reliability functions are easy to understand, do not require balanced data sets and the locations, slopes and shapes of the functions may be used to describe and to compare the across environment performance of test cultivars relative to a check. In addition, the functions are directly related to the stabilities of the test cultivar and the check. Our objectives are to use hybrid maize strip test trials to demonstrate how estimated reliability functions of test-check differences may be used to: (1) evaluate the similarity of responses of test cultivars relative to the check, (2) evaluate across environment variability of test cultivars and (3) demonstrate how reliability functions are related to the stability of the test and check cultivars.

Materials and methods

Estimating, graphing and testing reliability functions

If reliability functions of test-check differences are to be useful, it is necessary to estimate these functions based on field trial information. One approach is to obtain the fraction of environments in the trial where the test-check difference is greater than d and use this value as an estimate of $R_i(d)$. For example, if there are n environments where both the test and check cultivars are present, and there are X environments where the test-check difference is greater than d, then X/n is an unbiased estimate of $R_i(d)$. The computation of these estimates for all possible values of d results in an estimated reliability function. Confidence bounds may also be computed for the reliability function for any value of d (Nelson 1982).

Graphs of the estimated reliability functions are plotted as step functions (Fig. 1). These graphs may be used to visually describe how test-check differences vary over environments. Reliabilities (probabilities of outperforming the check) may also be simply obtained from the graph by identifying the value of the function when the test-check difference is 0 (Fig. 1).

If the decision maker is interested in statistically testing the equality of the reliability functions of two test cultivars, the Hollander bivariate symmetry test may be used on the assumption that the two test cultivars and the check are common in some set of representative environments (Hollander and Wolfe 1973; Kotz and Johnson 1983). In addition, the equality of the true reliabilities of several test cultivars may be tested using methods proposed by Cochran (1950) and Eskridge and Mumm (1992).

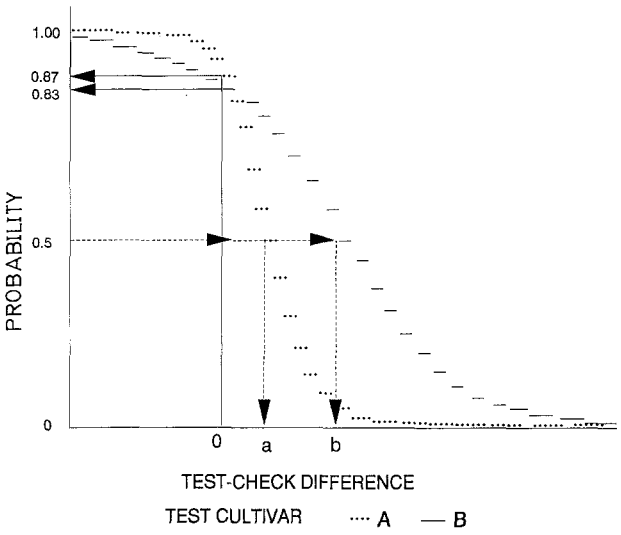

Fig. 1. Reliability functions of test-check differences for two hypothetical test cultivars illustrating reliabilities for test cultivars A and B (0.87, 0.83) and their medians of test-check differences (a and b)

Location, slope and shape of reliability functions

The more the reliability function is shifted to the right, the larger the mean difference and the more likely the test cultivar will outperform the check cultivar, assuming slope and shape are unchanged.

A steep reliability function indicates the test cultivar responds similarly to the check over environments. A gradually sloped function reflects more variability in performance differences and less similarity in response relative to the check. In Fig. 1, cultivar A has a smaller median difference (median of $A=a$; median of $B=b$), but has a response more similar to the check than cultivar B. Also, since cultivar B has a more gradual slope, B has a larger fraction of environments where it outperformed the check by a larger margin than what woud be expected if B simply had a larger mean performance.

In addition, the slope of the reliability function for a test cultivar is directly related to the stabilities of the test and check cultivars. If test-check differences for the ith test cultivar are normally distributed with mean μ_{di} and standard deviation σ_{di}, then the reliability function is

$$R_i(d) = P[Z > (d - \mu_{di})/\sigma_{di}] \quad (1)$$

where Z is a standard normal random variable and d is any value. The reliability function Eq. 1 is mathematically related to several commonly used stability parameters through σ_{di}. Finlay and Wilkinson's regression coefficient (β_i) (1963), Eberhart and Russell's deviation mean square ($\sigma^2_{\delta i}$) (1966) and Shukla's stability variance (σ^2_i) (1972) are functionally related to the variance of the test-check differences (σ^2_{di}) in the following ways:

Shukla: $\sigma^2_{di} = \sigma^2_i + \sigma^2_c$;

Eberhart-Russell: $\sigma^2_{di} = (\beta_i - \beta_c)^2 \sigma^2_I + \sigma^2_{\delta i} + \sigma^2_{\delta c}$; and

Finlay-Wilkinson: $\sigma^2_{di} = (\beta_i - \beta_c)^2 \sigma^2_I$,

where any parameter with subscript c is that parameter for the check and σ_I^2 = variance of the environmental index (Eskridge and Mumm 1992).

Equations for σ_{di}^2 expressed in terms of stability parameters may be substituted into Eq. 1 to demonstrate how the slope of the reliability function is directly related to these stability parameters. For two test cultivars with the same μ_{di}, the cultivar with the larger σ_{di}^2, as caused by a larger value of $|\beta_i - \beta_c|$ or $\sigma_{\delta i}^2$ or σ_i^2, will have a reliability function with a more gradual slope. Thus, a more gradual slope will indicate a less stable test cultivar. In Fig. 1, cultivar B is less stable than A.

The shape of the reliability function may also be useful in understanding what environmental factors either contribute or detract from the test cultivars response as compared to the check. For example, in Fig. 1, cultivar B has a larger chance of falling far below the check, as indicated by its longer tail on the left below cultivar A. Identification of environmental and management conditions that contributed to these poor responses relative to the check can aid in understanding where the cultivar has problems.

Figure 1 also demonstrates an important limitation of basing selection solely on reliability (probability of outperforming the check). The two test cultivars have nearly identical reliabilities, but most decision makers would consider cultivar B to be clearly superior since it outperforms the check by a larger amount than A, as indicated by most of its reliability function being to the right of cultivar A.

On-farm hybrid maize trials

To demonstrate the approach, data were taken from on-farm strip test trials grown cooperatively by Pioneer Hi-bred International and farmers at 122 environments in Iowa, Illinois, and Nebraska during 1988, 1989 and 1990. The strip trials were single replicate on-farm tests with from 4 to 12 hybrids grown in maturity order. The set of hybrids grown varied from environment to environment, and plots were chosen where the 4 hybrids used in this analysis were grown together. Hybrid 4 was generally well-adapted to the environments in this study. Plots at the individual sites varied from 4 to 8 rows wide and from 91 to 229 m in length. Spacing between the rows varied from 76 to 91 cm. The plots varied from about 0.04 to 0.1 ha. The plots were planted and harvested using cooperators' equipment. Grain yields were determined using weight wagons provided by Pioneer agronomists. The plots were planted to stand and were not gleaned during harvest. All grain weights were adjusted to 15.5% grain moisture.

Using yields from these trials and hybrid 4 as the check, we estimated reliability functions of the test-check differences for the 3 test hybrids. These functions were used to graphically describe how the differences varied over environments, identify environments where hybrids performed poorly and estimate the reliabilities of the test hybrids. In addition, equality of the test cultivars' true reliability functions was tested using the Hollander bivariate symmetry test. Equality of the test cultivars' reliabilities was tested using Cochran's Q test (Cochran 1950). Finlay and Wilkinson's regression coefficient (β_i) (1963), Eberhart and Russell's deviation mean square ($\sigma_{\delta i}^2$) (1966) and Shukla's stability variance (σ_i^2) (1972) were also estimated for each of the 4 cultivars.

Results

Reliability functions were obtained for each of the 4 maize hybrids from the strip test information based on all environments where all 4 hybrids were present. Means, mean differences from the check and their standard deviations, stability statistics and reliabilities were obtained for each of the 4 maize hybrids (Table 1). The hybrids' reliability functions clearly illustrated how the test-check differences varied over environments (Fig. 2). The Hollander bivariate symmetry test failed to reject the hypothesis of coincidence of reliability functions of hybrids 1 and 3 ($P \approx 0.40$), but the reliability functions of hybrids 1 and 3 differed from the reliability function of hybrid 2 ($P < 0.001$). The most obvious feature of the three reliability functions was that hybrid 2 was shifted to the right and had a more gradual slope as compared with test hybrids 1 and 3. The shift to the right indicated that hybrid 2 had higher mean performance.

Hybrid 2's more gradual slope indicated that it had a larger fraction of environments where it outperformed the check by a larger margin than what would be expected if hybrid 2 had only a mean performance larger than either hybrid 1 or 3. Having a performance beyond what would be expected due to simply a larger mean apparently meant that hybrid 2 had an increased ability to take advantage of environmental conditions where it outperformed the check. Hybrid 2's more gradual slope also indicated it was less stable than hybrids 1 or 3. This larger instability was reflected through larger values of $|b_i - b_c|$, $S_{\delta i}^2$ and $\hat{\sigma}_i^2$.

In environments where hybrid 2's yield was more than 0.5 ton ha^{-1} below the check, its performance decreased faster than would have been expected as indicated by the 'notch' in the left tail of hybrid 2's reliability function. The identification of environmental conditions which contributed to hybrid 2's poor performance rela-

Table 1. Means, yield differences standard deviations, stability statistics and reliabilities of 4 maize hybrids over 122 common environments

Hybrid[a]	Yield mean[b]	Test-check differences[b]		Stability statistics[c]			$R_i(0)$[d]
		Mean	SD	b_i	$S_{\delta i}^2$	$\hat{\sigma}_i^2$	
1[A]	9.90	0.04	0.79	0.93	0.20	0.23	0.53[A]
2[B]	10.69	0.83	1.01	1.10	0.47	0.81	0.77[B]
3[A]	9.99	0.13	0.75	0.97	0.18	0.17	0.56[A]
4 (check)	9.86	–	–	1.00	0.29	0.37	–

[a] Hybrid reliability functions with common letters do not differ ($\alpha = 0.05$) based on Hollander bivariate symmetry test

[b] Metric tons ha^{-1}

[c] b_i, Finlay and Wilkinson's regression coefficient (1963); $S_{\delta i}^2$, Eberhart and Russell's deviation mean square (1966), $\hat{\sigma}_i^2$, Shukla's stability variance (1972)

[d] The estimated chance that the test hybrid outperforms the check in a future environment (reliability). Hybrid reliabilities with common letters do not differ ($\alpha = 0.05$) based on Cochran's Q test (Cochran 1950)

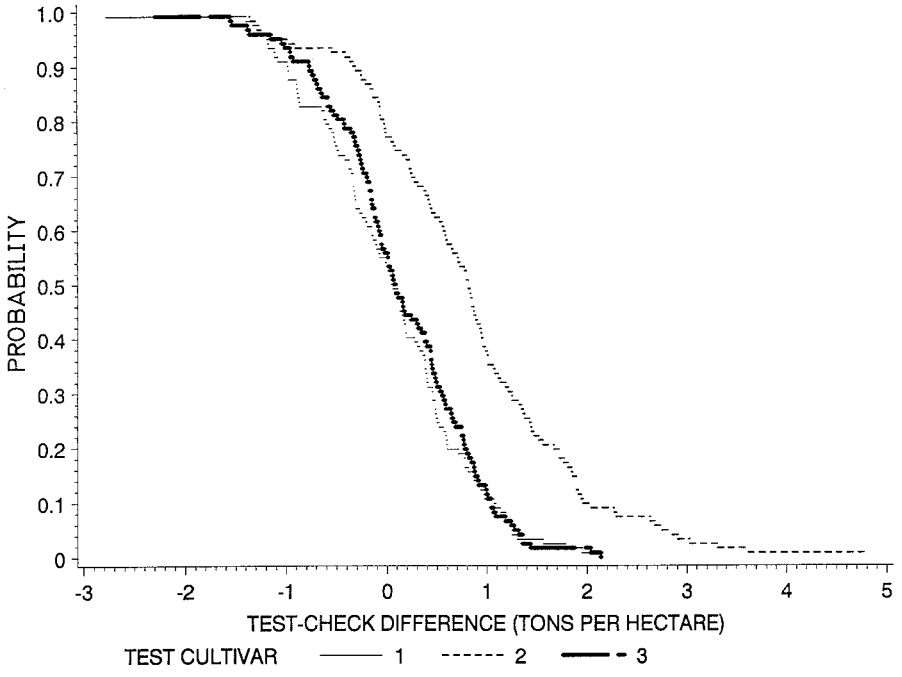

Fig. 2. Reliability functions of test-check differences for three Pioneer corn hybrids based on data from on-farm trials conducted in the years 1988 through 1990

tive to the check would have been useful in determining where the hybrid had performance problems. However, given the information available on these locations, it was not possible to identify why hybrid 2 performed so poorly relative to the check in these environments, and the notch was considered to be caused by unexplained variation associated with strips in the plots.

If selections among these hybrids were to be based on yield alone, the reliability functions demonstrated how hybrid 2 should be preferred to hybrids 1 and 3 by all decision makers since it had a larger probability of outperforming the check by more than d bushels for all values of d. This superior performance was indicated by hybrid 2's reliability function never falling to the left of either hybrid 1 or 3. In addition, Cochran's Q test (Cochran 1950) indicated that the reliabilities of hybrids 1 and 3 did not differ significantly ($P > 0.50$), while the reliability of hybrid 2 differed significantly from hybrids 1 and 3 ($P < 0.02$) (Table 1).

Discussion

The usefulness of reliability functions of test-check differences in identifying superior cultivars is based on several assumptions. First, decision makers using the approach are assumed to be primarily interested in the performance of test cultivars relative to a check. If a decision maker values cultivars using other approaches, for example, an index of economic worth, more direct methods are available. Second, the approach is based on the assumption that the trial information is obtained from environments that are representative of those environments where the check is well-adapted. If the check is planted outside its range of adaptability, its performance may fall well below a test cultivar whereas in its range of adaptability, it may outperform the test cultivar. Thus, the comparison of a test cultivar with a check planted in environments to which it is not adapted may give reliability functions that are biased against the check.

There are several advantages to using reliability functions of test-check differences in comparing and identifying superior test cultivars. The approach is conceptually straightforward and can be illustrated using simple graphs, thereby allowing easy understanding for many different types of decision makers be they breeders, growers or management. Moreover, because the method is pairwise in nature it can maximize the number of environments for a comparison between a test cultivar and a check, thus avoiding problems of imbalance that occur when differing sets of cultivars are included in different trials (Bradley et al. 1988).

Reliability functions of test-check differences also provide a basis for comparing test cultivars over a range of environments. The location, slope and shape of the functions give a general indication of how the differences are distributed over environments. Steeper slopes indicate less variability and more similarity in response relative to the check. Functions shifted more to the right imply larger performance differences compared to the check, while the shapes of reliability functions may be useful in identifying environments where a cultivar has performance problems relative to the check. Additional information on these environments could be used to iden-

tify environmental and management factors that contribute to such problems.

Reliability functions of test-check differences are also useful in comparing the stabilities of several test cultivars. Commonly used stability statistics b_i (Finlay and Wilkinson 1963), $S^2_{\delta i}$ (Eberhart and Russell 1966) and $\hat{\sigma}^2_i$ (Shukla 1972) are related to a test cultivar's reliability function. These stability statistics (b_i, $S^2_{\delta i}$, $\hat{\sigma}^2_i$), however, can only be used as relative measures since each depends on the particular set of cultivars being evaluated (Lin et al. 1986). Reliability functions have a broader inference base since they only depend on the check and the particular test cultivar being considered and they do not depend on other cultivars.

The approach is also appealing because of its generality. The use of reliability functions of test-check differences requires no assumption about the form of the probability distribution of test-check differences. The use of reliability functions has been well documented in the quality engineering literature (Nelson 1982).

There are several consequences that need to be considered when using reliability functions of test-check differences to identify superior plant cultivars. Because the method is based on a single trait, it fails to consider compromises among traits that are necessary when making final selections. However, the method could be applied to multiple trait selection indices or extended for use with multivariate vectors of traits. Because the approach compares test cultivars with a common check, the choice of a check can have a major impact on the estimated reliability functions. In situations where comparisons with several checks are desired, it may be necessary to compute different reliability functions using different checks. A moderate number of environments (say 15 or more) also are generally required to estimate points on the reliability function with an acceptable level of precision. For example, when responses are normally distributed and the true reliability is 0.85, 16 environments are needed to be within 0.15 of the true value. Finally, reliability functions can be useful aids to plant breeders when evaluating cultivars in the presence of GE interaction, but breeders should not use the approach in lieu of understanding the biological nature of these interactions.

Acknowledgements. We would like to thank the numerous cooperators for helping obtain the data, the reviewers for their helpful comments and Leona Barratt for help with making the figures.

References

Bradley JP, Knittle KH, Troyer AF (1988) Statistical methods in seed corn product selection. J Prod Agric 1:34–38

Cochran WG (1950) The comparison of percentages in matched samples. Biometrika 37:256–266

Eberhart SA, Russell WA (1966) Stability parameters for comparing varieties. Crop Sci 6:36–40

Eskridge KM, Mumm RF (1992) Choosing plant cultivars based on the probability of outperforming a check. Theor Appl Genet 84:494–500

Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant-breeding programme. Aust J Agric Res 14:742–754

Hollander M, Wolfe DA (1973) Nonparametric statistical methods. Wiley, New York

Jones TA (1988) A probability method for comparing varieties against checks. Crop Sci 28:907–912

Kotz S, Johnson NL (1983) Encyclopedia of statistical sciences. Wiley, New York

Lin CS, Binns MR, Lefkovitch LP (1986) Stability analysis: where do we stand? Crop Sci 26:894–900

Nelson W (1982) Applied life data analysis. Wiley, New York

Shukla GK (1972) Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29:237–245