Theor Appl Genet (1992) 84:494–500

THEORETICAL
AND
APPLIED GENETICS

Choosing plant cultivars based on the probability of outperforming a check *

K. M. Eskridge and R. F. Mumm

Department of Biometry, University of Nebraska, Lincoln, NE 68583-0712, USA

Received August 15, 1991; Accepted December 19, 1991
Communicated by A. R. Hallauer

Summary. A major consideration in most plant breeding programs is the development of cultivars that have high probabilities of outperforming the check cultivar in a broad range of environments. Methods are presented for estimating and testing hypotheses regarding these probabilities, which are termed reliabilities. Reliabilities are shown to be directly related to several commonly used stability parameters. Data from international maize yield trials are used to illustrate and evaluate the repeatability of the approach. Results indicate that reliabilities can be useful aids to plant breeders since they (1) are easy to understand and compute, (2) are indices that weigh the importance of the difference in performance relative to stability, and (3) are potentially useful as genetic parameters since they are generally repeatable across randomly sampled sets of environments.

Key words: Selection – Stability analysis – Performance testing – Reliability

Introduction

A major consideration in most plant breeding programs is the development and identification of new cultivars that perform better than accepted cultivars over a broad range of environments. In response to this concern, breeding programs have increasingly emphasized "head-to-head" or pairwise cultivar comparisons as a means of comparing test cultivars to checks across a wide range of environments (Bradley et al. 1988; Jones 1988). Such head-to-head comparisons generally involve compilation of information on the test cultivar and the check over all environments where both are grown. These data are then used to estimate and test for true mean differences between the two cultivars for a number of different traits. Such pairwise comparisons maximize the number of locations in the comparisons, avoid the problem of unbalancedness that occurs when differing sets of cultivars are included in different tests, and are useful for making comparisons in the presence of genotype-environment interaction (Bradley et al. 1988; Jones 1988).

Although the pairwise approach to cultivar evaluation has many advantages, the specific methods of analysis used in conjuction with this approach may not provide the information most relevant to the plant breeder. The overriding concern of many breeders is the identification of test cultivars that have a high *probability* of outperforming the check in environments where the check is normally grown. Statements commonly made about pairwise cultivar comparisons usually apply to 'true' mean trait values over a 'population' of environments (Bradley et al. 1988; Jones 1988). Knowing the true mean values for each cultivar does not provide much insight into the chances of the test cultivar outperforming the check. In addition, a 'true' mean is an abstract concept which may not be clear to people other than plant breeders (e.g., growers, sales managers, etc.). The usefulness of pairwise cultivar evaluation could be enhanced by developing a decision-making tool that (1) quantifies the *probability* that a test cultivar outperforms the check over a broad range of environments and (2) is easily interpretable to all decision makers – be they breeders, growers, or other interested parties.

One group of decision-making models that is easy to understand and may be used to quantify the chances that a cultivar will outperform the check is based on the assumption of **safety-first** behavior (Eskridge 1990a, b;

* Contribution from the Agricultural Research Division, University of Nebraska, Lincoln, Neb., Paper no. 9653
Correspondence to: K.M. Eskridge

1991; Eskridge et al. 1991). Stated in terms of selection, a breeder practices safety-first behavior if he/she is primarily concerned with the probability of achieving (or failing to achieve) an acceptable response for each cultivar being considered. Safety-first decision models may be used to quantify the head-to-head cultivar comparison perspective by assuming the breeder is primarily concerned with the probability that a new cultivar outperforms the check cultivar in environments where the check cultivar is well suited. More concisely, the decision maker is primarily concerned with:

$$P(Y_i - Y_c > 0) \qquad (1)$$

where Y_i and Y_c are responses of the ith cultivar and the check cultivar, respectively, and P denotes probability. Equation (1) is defined as the reliability of the ith test cultivar where a reliable cultivar has a high probability of outperforming the check. Reliability is a commonly used concept in machine life testing and quality engineering (Nelson 1982).

Reliability of test cultivars (Eq. (1)) can be useful to breeders in identifying superior cultivars for several reasons. The reliability of a cultivar is an easily understood measure of "riskiness" of a cultivar, assuming the major risk facing the breeder is choosing a test cultivar that fails to outperform the check. Choice of a cultivar with a reliability near 0.5 is risky since, on the average, the test cultivar fails to outperform the check in 50% of the environments. Another test cultivar with a reliability of 0.9 is much less risky since it would be expected to fall short of the check in only 10% of the environments.

In addition, reliability is directly related to several commonly used measures of stability. As a result, reliability is an index that explicitly weighs the importance of differences in performance relative to stability, when comparing the test cultivar to the check. Most univariate methods used by plant breeders to identify stable, high-performing cultivars are limited since they do not clearly specify how to weigh the importance of performance relative to stability (Finlay and Wilkinson 1963; Eberhart and Russell 1966; Shukla 1972). Also, reliability is more general than traditional methods that make statements about true means. Reliability will provide similar conclusions on cultivar preferences when cultivars differ little in stability, but it can result in cultivar preferences that are quite different from those based on traditional methods when stabilities differ substantially. Furthermore, the reliability explicitly incorporates genotype x environment interaction since it involves the difference of test and check cultivars across environments.

The objectives of this paper are to use international maize variety trials to (1) illustrate how the reliability of a cultivar (Eq. (1)) may be used to aid decision makers with choosing among test cultivars, (2) demonstrate how reliability is related to several stability measures and how it weighs performance relative to stability, and (3) evaluate the usefulness of reliability by comparing the repeatability of the approach with the repeatability of some commonly used stability measures.

Materials and methods

Estimating and testing reliabilities

If reliability is to be useful in aiding the breeder with identifying superior cultivars, it is necessary to use field trial information to estimate and test hypotheses regarding these reliabilities. Reliabilities may be estimated and tested using at least two different approaches.

Normally distributed differences. Let $d_i = Y_i - Y_c$ be the difference between the response of the ith test cultivar and that of the check. If d_i is normally distributed over the population of environments with mean μ_{di} and standard deviation σ_{di} then the reliability (Eq. (1)) of the ith test cultivar may be stated as:

$$P(Z > -\mu_{di}/\sigma_{di}) \qquad (2)$$

where Z is a standard normal random variable. μ_{di} and σ_{di} are not known and can be estimated using the sample mean difference ($\bar{y}_{di}$) and standard deviation (s_{di}) based on field trial information. These values may be substituted for μ_{di} and σ_{di} in Eq. (2) to estimate reliability (Nelson 1982). Reliability for the ith cultivar estimated in this way will be denoted RN_i. (The *t* distribution could be used to estimate reliabilities, but such estimates would differ only slightly from those based on the standard normal distribution when the trial has enough environments to provide fairly precise reliability estimates. See discussion.) Also, approximate $100(1-\alpha)\%$ confidence intervals for the reliability of a test cultivar may be estimated (Nelson 1982). Additionally, when the check and test cultivars under consideration are common in some set of environments, the Wald test may be used to test equality of the reliabilities and contrasts among reliabilities for several test cultivars simultaneously (see Appendix).

Nonparametric approach. The assumption of normality may not be justified with some traits. An alternative approach may be used to estimate the reliability of the ith cultivar without making any assumptions about how the differences (d_i's) are distributed. With this approach, the sample proportion of environments where the test cultivar outperformed the check is an estimate of the reliability for the test cultivar. Reliabilities estimated in this way will be denoted R_i. Confidence intervals for proportions (Steel and Torrie 1980) may be used to obtain interval estimates of the reliability of a test cultivar. Additionally, when the check and test cultivars under consideration are all present in some set of environments, Cochran's Q test (Cochran 1950) may be used to test equality of reliabilities and contrasts among reliabilities for several test cultivars simultaneously. The idea of this test is similar to an *F* test for treatment in a randomized complete block experiment with environments as blocks, test cultivars as treatments, and where the responses for each environment-cultivar cell is 0 if $d_i \leq 0$ and 1 if $d_i > 0$ (Winer 1971).

Relationship between reliability and several stability measures

Reliability can be shown to be directly related to several commonly used stability measures when reliability is defined with Eq. (2). 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 cultivar check differences (σ^2_{di}) in the following ways (see Appendix):

Shukla: $\sigma_{di}^2 = \sigma_i^2 + \sigma_c^2$
Eberhart-Russell: $\sigma_{di}^2 = (\beta_i - \beta_c)^2\, \sigma_I^2 + \sigma_{\delta i}^2 + \sigma_{\delta c}^2$
Finlay-Wilkinson: $\sigma_{di}^2 = (\beta_i - \beta_c)^2\, \sigma_I^2$

where any parameter with subscript c is that parameter for the check, and σ_I^2 = variance of the environmental index.

Equations of σ_{di}^2 expressed in terms of stability parameters may be substituted into Eq. (2) to demonstrate how reliability explicitly weighs the importance of the difference in performance (μ_{di}) relative to stability. In general, for a given positive difference in mean performance, stability parameters that result in larger values of σ_{di}^2 will reduce the reliability of the test cultivar.

By using the different definitions of stability in Eq. (2), it can be seen how reliability is related to these particular stability parameters when the mean difference (μ_{di}) is positive. For Shukla's model, in which the mean difference (μ_{di}) and other terms, are held constant, as Shukla's variance for the test cultivar (σ_i^2) becomes larger, the ratio μ_{di}/σ_{di} becomes smaller and the reliability of the cultivar is reduced. This result is reasonable since the larger Shukla's variance the less stable and more unreliable the cultivar. For Eberhart and Russell's model, in which other terms are held constant, an increase in the ith cultivar's mean square deviation ($\sigma_{\delta i}^2$) also reduces the reliability of the cultivar. Again, this result is reasonable since larger mean square deviations imply less stability and thus less reliability. Similarly, for Finlay and Wilkinson's approach, in which all other terms are held constant, the larger the absolute difference between slope coefficients of the test cultivar and the check ($\beta_i - \beta_c$), the smaller the reliability.

The above reasoning is useful to see how reliability is related to these stability parameters. However, holding μ_{di} and 'other terms' constant as the stability parameters of interest are varied is not possible in application. To empirically assess the relationship between reliability and these stability measures with regard to ranking cultivars, rank correlations among reliabilities and the various measures were computed.

Comparing the repeatability of reliabilities with other measures

For the reliability of a cultivar to be useful to the plant breeder, it is necessary that this value actually be representative of the genetic characteristics of the cultivar under consideration. If a parameter estimate is truly a measure of the genetic features of a cultivar, then the ranking of a group of cultivars based on the estimated parameter should be fairly consistent between any two different sets of environments randomly sampled from the population of environments under consideration. In order to assess the repeatability of reliability as a measure of genetic characteristics compared to other measures, for each of four experimental variety trials (EVTs 12, 13, 14A and 14B), environments were randomly separated into two sets, and calculations were made for each set. (EVTs 16A and 16B were not used due to their limited number of environments.) This process was repeated six different times (runs) to ensure dependable results. Mean yields, reliabilities (R_i, RN_i), Finlay and Wilkinson's (1963) regression coefficients (b_i), Eberhart and Russell's (1966) mean square deviations ($S_{\delta i}^2$), and Shukla's (1972) stability variances ($\hat{\sigma}_i^2$) were estimated for each set of environments and each run. For each statistic, rank correlations across varieties were then computed. Large rank correlations indicated that an estimated parameter consistently ranked varieties over the two sets of environments, thus indicating that it was repeatable and a useful measure of genetic features of a variety.

International maize variety trials

Six CIMMYT experimental variety yield trials (EVTs) were used to illustrate how reliabilities could be used to aid with choosing among cultivars and to compare the repeatability of reliability with other measures of stability. Description of the trials were given in CIMMYT (1990) and are summarized in Table 1. Locations were included in the analyses only if the coefficient of variation of yield was less than 30% and the coefficient of variation of plants harvested was less than 20%. Since local checks differed for each location, local checks were not included in the analyses. For each EVT, the check was identified as the oldest reference entry with the lowest overall mean yield. These reference entries were considered adequate check varieties because most were older varieties found in past trials to be fairly well adapted to the environmental conditions under which the trials were grown. Given each entry's sample mean difference ($\bar{y}_{di}$), standard deviation (s_{di}), and the number of environments where each entry outperformed the check, reliabilities were computed using the normal and nonparametric approaches (RN_i and R_i) as outlined above.

Table 1. Description, check variety, number of varieties, number of environments analyzed, and yield for six CIMMYT 1988 Experimental Variety Trials (EVT's) (CIMMYT 1990)

EVT	Description of varieties	Check variety	Number of varieties[a]	Number of environments	Yield	
					Mean mg/ha	Range mg/ha
12	Tropical lowland, late maturity, white grain, developed from four populations	Across 7729	10	47	4.97	1.25– 9.87
13	Tropical lowland, late maturity, yellow grain, developed from four populations	Across 7627	15	58	5.36	1.73– 8.84
14A	Tropical lowland, early and intermediate maturity, yellow grain, developed from two populations	Across 8331	17	63	4.75	1.38– 8.08
14B	Tropical lowland, early and intermediate maturity, white grain, developed from four populations	Across 7823	19	43	4.24	1.22– 7.77
16A	Tropical mid-altitude, early and intermediate maturity, yellow grain, developed from four populations	Across 7748	15	24	4.40	0.57–10.50
16B	Tropical mid-altitude, intermediate and late maturity, white grain, developed from four populations	Across 7734	16	22	5.14	0.93–11.24

[a] Excluding local checks

Table 2. Mean yields, number of environments where variety outperforms check (n), mean differences ($\bar{y}_{di}$), standard deviations of differences (s_{di}), reliabilities (R_i and RN_i), Finlay and Wilkinson's regressions coefficient (b_i), Eberhart and Russell's deviation mean square ($S^2_{\delta i}$), and Shukla's stability variance (σ^2_i) for maize varieties in 1988 CIMMYT EVT 13 with Across 7627 RE as check and 58 environments

Variety number	Mean yield (mg/ha)	n	$\bar{y}_{di}$	s_{di}	R_i[a]	RN_i[b]	b_i	$S^2_{\delta i}$	$\hat{\sigma}^2_i$
3	5.666	55	0.850	0.560	0.948	0.935	0.968	0.141	0.145
7	5.642	52	0.826	0.700	0.897	0.881	0.977	0.204	0.216
5	5.578	51	0.762	0.703	0.879	0.861	1.032	0.227	0.243
6	5.706	50	0.890	0.769	0.862	0.877	0.994	0.297	0.320
4	5.481	49	0.665	0.638	0.845	0.851	0.970	0.195	0.206
13	5.453	49	0.637	0.719	0.845	0.812	1.050	0.169	0.180
15	5.573	48	0.797	0.692	0.828	0.863	1.049	0.197	0.211
2	5.372	45	0.556	0.672	0.776	0.796	1.012	0.197	0.207
10	5.237	44	0.421	0.621	0.759	0.751	1.053	0.229	0.248
9	5.285	43	0.469	0.557	0.741	0.800	1.024	0.164	0.171
11	5.300	43	0.484	0.652	0.741	0.771	1.002	0.190	0.199
1	5.058	41	0.242	0.563	0.707	0.666	0.982	0.188	0.197
12	5.140	41	0.324	0.610	0.707	0.703	0.943	0.170	0.182
8	5.121	40	0.305	0.673	0.690	0.675	1.041	0.288	0.313
14	4.816	—	—	—	—	—	0.901	0.184	0.212

[a] Cochran's Q = 36.72, P = 0.00045 for hypothesis of no difference among true reliabilities
[b] Wald Statistic = 51.69, $P = 1.5 \times 10^{-6}$ for hypothesis of no difference among true reliabilites

Table 3. Rank correlations between mean yield (mean), reliability (R_i and RN_i), regression coefficient (b_i), mean square deviation ($S^2_{\delta i}$), and stability variance ($\hat{\sigma}^2_i$) for 1988 CIMMYT EVTs

EVT	Measure	R_i	RN_i	b_i	$S^2_{\delta i}$	$\hat{\sigma}^2_i$
12	Mean	0.743	0.944	0.333	0.111	−0.067
12	R_i	1.000	0.743	0.114	0.000	0.057
12	RN_i		1.000	0.167	0.000	−0.056
13	Mean	0.805	0.890	0.010	0.162	0.086
13	R_i	1.000	0.805	−0.112	0.022	0.045
13	RN_i		1.000	−0.121	0.011	0.033
14A	Mean	0.731	0.883	0.103	−0.118	−0.088
14A	R_i	1.000	0.696	−0.017	−0.087	−0.070
14A	RN_i		1.000	0.000	−0.100	−0.083
14B	Mean	0.711	0.961	0.661	0.322	0.275
14B	R_i	1.000	0.670	0.573	0.380	0.325
14B	RN_i		1.000	0.647	0.294	0.255
16A	Mean	0.702	0.758	0.562	0.524	0.505
16A	R_i	1.000	0.794	0.449	0.633	0.633
16A	RN_i		1.000	0.648	0.736	0.714
16B	Mean	0.853	0.905	0.467	0.233	0.200
16B	R_i	1.000	0.834	0.441	0.167	0.147
16B	RN_i		1.000	0.352	0.086	0.067

Results

To illustrate the use of reliability as an aid to choosing among test cultivars, both nonparametric and normal reliabilities (R_i and RN_i) for EVT 13 were computed for all test varieties using 'Across 7627 RE' as the check variety (Table 2). Means, standard deviations, and standard stability statistics were included for comparison purposes. Estimated reliabilities for all varieties were larger than 0.65, meaning that all varieties had better than a 65% chance of outperforming the check. Reliabilities differed substantially among varieties as depicted by large values of Cochran's Q (Q = 36.72, P = 0.000458) and the Wald statistic (W = 51.69, $P = 1.5 \times 10^{-6}$). Both nonparametric and normal reliabilities (R_i and RN_i) resulted in similar estimated reliability values for each variety, revealing that either measure could be used with confidence with these data. In addition, a high mean yield did not necessarily imply that a variety would be highly reliable. For example, variety 6 had the largest mean yield, but was found to have the fourth and third highest reliabilities for R_i and RN_i, respectively.

The relationships between reliability estimates (R_i and RN_i) and the three stability statistics (b_i, $S^2_{\delta i}$, $\hat{\sigma}^2_i$) were quantified for all EVTs using Kendall's Tau rank correlations (Table 3). Reliabilities were strongly correlated with the mean yield ($r > 0.70$) in all EVTs, indicating that varieties with large means tend to have larger reliabilities. R_i and RN_i were somewhat positively correlated ($r > 0.35$) with b_i in EVTs 14B, 16A, and 16B. This result appeared to be caused by b_i being somewhat correlated with the mean ($r > 0.46$) which, in turn, was highly correlated with reliability. Deviation mean squares ($S^2_{\delta i}$) and Shukla's stability variance ($\hat{\sigma}^2_i$) were poorly correlated ($r < 0.39$) with both reliability estimates (R_i, RN_i) in five of six EVTs. These poor correlations indicated that both deviation mean squares and Shukla's stability variance had an insignificant impact on reliabilities. In EVT 16A, deviation mean squares and Shukla's variance were positively correlated ($r > 0.63$) with both reliabilities. For this EVT, large correlations were likely caused by two distinct groups of varieties, one group with large means, large

Table 4. Rank correlations between two randomly separated sets of environments based on the mean yield, reliabilities (R_i and RN_i), regression coefficient (b_i) mean square deviation ($S^2_{\delta i}$), and stability variance ($\hat{\sigma}^2_i$) for four 1988 CIMMYT EVTs when environments were randomly separated in six different runs

EVT	Number of varieties	Run	Rank correlations					
			Mean yield$_i$	R_i	RN_i	b_i	$S^2_{\delta i}$	$\hat{\sigma}^2_i$
12	10	1	0.511	0.561	0.222	0.689	0.022	0.156
		2	0.467	0.159	0.333	0.156	0.289	0.200
		3	0.244	0.223	0.222	0.600	0.067	0.200
		4	0.511	0.381	0.556	0.200	−0.022	0.244
		5	0.289	0.571	0.333	0.600	0.200	0.244
		6	0.022	−0.125	−0.222	0.511	−0.111	0.200
		Mean	0.341	0.295	0.251	0.459	0.074	0.207
13	15	1	0.752	0.485	0.604	−0.181	0.219	0.086
		2	0.752	0.281	0.560	0.105	−0.162	−0.238
		3	0.810	0.417	0.604	0.429	−0.181	−0.200
		4	0.771	0.330	0.406	0.048	−0.505	−0.467
		5	0.790	0.694	0.626	0.048	−0.219	−0.200
		6	0.638	0.339	0.451	0.295	0.010	0.010
		Mean	0.752	0.424	0.542	0.124	−0.140	−0.168
14A	17	1	0.588	0.356	0.500	0.103	−0.029	−0.015
		2	0.338	0.423	0.283	0.308	−0.103	−0.103
		3	0.206	0.056	0.283	0.029	−0.059	−0.074
		4	0.323	0.212	0.267	0.191	−0.059	0.000
		5	0.529	0.247	0.350	0.103	−0.176	−0.147
		6	0.559	0.351	0.550	0.088	0.015	0.073
		Mean	0.424	0.274	0.372	0.137	−0.068	−0.044
14B	19	1	0.485	0.528	0.516	0.427	−0.017	0.053
		2	0.497	0.171	0.516	0.474	0.252	0.287
		3	0.532	0.472	0.386	0.321	0.123	0.146
		4	0.450	0.446	0.464	0.275	0.111	0.181
		5	0.567	0.424	0.647	0.263	0.216	0.158
		6	0.661	0.504	0.608	0.649	0.356	0.427
		Mean	0.532	0.424	0.523	0.401	0.173	0.209

deviation mean squares, and large values of Shukla's stability statistic, while the other group had relatively small values of means, deviation mean squares, and Shukla's variances.

To assess the repeatability of reliabilities compared to other statistics, rank correlations of the mean yield, reliabilities, and stability statistics were computed between two randomly chosen sets of enviroments for each run (Table 4). The sizes of rank correlations differed substantially depending on the EVT and statistic. For all EVTs, the rank correlations for both reliabilities (R_i and RN_i) were somewhat lower than the rank correlation for the mean yield, meaning that both reliability measures were not as repeatable as the mean. However, in most runs both reliabilities had rank correlations that were larger than rank correlations of the joint regressions coefficient (b_i) and substantially larger than rank correlations of the mean square deviation ($S^2_{\delta i}$) and the stability variance ($\hat{\sigma}^2_i$). This result demonstrated that reliabilities, although not as repeatable as the mean, had a considerably better repeatability than several of the most commonly used stability statistics. In addition, the repeatabilities of Eberhart and Russell's mean square deviation ($S^2_{\delta i}$) and Shukla's stability variance ($\hat{\sigma}^2_i$) were generally quite small ($r < 0.2$), indicating that neither appeared to measure the genetic features of a cultivar.

Discussion

The usefulness of reliability as a decision aid for identifying superior test cultivars is based on two assumptions. First, the breeder is principally concerned with identifying test cultivars that have a good chance of outperforming the check in environments where the check is normally grown. If the breeder is primarily interested in selecting test cultivars based on other criteria, such as the largest mean or highest stability, then more direct methods are available. A second assumption is that the test trials are conducted in environments that are representative of the population of environments where the check cultivar is well adapted. Using a check cultivar that is specifically adapted to only certain environments can give unrealistic reliability estimates. Such a check planted outside its range of adaptability may fall well below a test cultivar in performance, whereas within the check's range of adaptation, it may always be superior.

Use of reliability as a decision tool has several advantages. Reliability is conceptually straightforward and is based on the reasonable assumption that the breeder is primarily interested in identifying test cultivars that have a high probability of outperforming the check. Given this perspective, reliability can be directly interpreted as the riskiness of a test cultivar. Also, reliability is more general than traditional tests and confidence intervals on true means since reliability will provide similar conclusions on cultivar preferences when cultivars differ little in stability. But when stabilities differ substantially, reliability can result in cultivar preferences that are quite different from those based on traditional methods. Moreover, the use of reliability does not depend on the same set of cultivars being grown in all environments. Any location or year where the test cultivar and check are grown within a reasonable proximity of one another will provide additional information on the reliability. Data may also be obtained from other tests conducted by other breeding programs. Use of pedigrees which are highly related to the test cultivar may also be substituted in lieu of the test cultivar if additional information is required. However, such results become less applicable as the similarity between the test cultivar and its substitute decreases.

Also, reliability is functionally related to several commonly used stability statistics (b_i, $S_{\delta i}^2$, $\hat{\sigma}_i^2$). However, reliability estimates are more repeatable than these stability statistics and thus are better measures, of the genetic characteristics of cultivars. In addition, these stability statistics (b_i, $S_{\delta i}^2$, $\hat{\sigma}_i^2$) can only be used as relative measures since each depends on the particular set of cultivars being evaluated (Lin et al. 1986). Reliability of a test cultivar has a broader inference base than b_i, $S_{\delta i}^2$, or $\hat{\sigma}_i^2$ because it only depends on the check and the particular test cultivar being considered and does not depend on other test cultivars in the trial.

Another advantage of reliability is that it is an index that explicitly weighs the importance of the difference in performance relative to stability. This property relieves the plant breeder from having to make decisions about how to weigh the importance of performance to stability when making final selections. Viewed as an index that explicitly combines both performance and stability through the ratio μ_{di}/s_{di} in Eq. (2), reliability compares favorably with other indices that combine performance and stability. Reliability has fewer assumptions and is easier to understand than the expected utility stability indices proposed by Barah et al. (1981) and Eskridge and Johnson (1991). Reliability does not require special 'disaster' parameters as do safety-first stability indices (Eskridge 1990a; Eskridge 1991; Eskridge et al. 1991). In addition, reliability can be used to obtain a complete ranking of the cultivars under test in contrast to stochastic dominance as used by Menz (1980) to categorize wheat varieties as risk-efficient or risk-inefficient.

Also, reliabilities can be estimated fairly precisely with a moderate number of environments. When responses are normally distributed and the true reliability is 0.85, about 16 environments are required to be within 0.15 of the true value. A further advantage of reliability is that statistical procedures for estimating and testing reliabilities have been well developed in the statistical reliability literature (Nelson 1982).

However, there are several limitations of using reliability to identify superior plant cultivars. Because the procedure compares test cultivars with a common check, the choice of a check can have a major impact on the reliabilities. The use of a specifically adapted check in trials conducted over a wide range of environments can result in unrealistic reliabilities. In this study CIMMYT reference entries were considered to be adequate checks since most were found to be fairly broadly adapted based on information from past trials. Also, in situations where there are several checks and it is not clear which is the most appropriate, it may be necessary to compute reliabilities using several different checks. Reliability can be a useful aid to plant breeders when selecting cultivars in the presence of genotype x environment interaction, however breeders should not use reliability in lieu of understanding the biological nature of genotype x environment interaction. Also, a larger number of environments will be needed to precisely estimate reliabilities when compared to other statistics such as the mean. Finally, more research is needed to assess the impact of economic factors, such as production costs, on the approach.

Acknowledgements. We would like to thank Dr. Pat Byrne, CIMMYT, and the numerous cooperators in the national research programs for providing the data used in this investigation.

References

Anderson TW (1958) An introduction to multivariate statistical analysis. Wiley, New York

Barah BC, Binswanger HP, Rana BS, Rao NGP (1981) The use of risk aversion in plant breeding: concept and application. Euphytica 30:451–458

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 WR (1966) Stability parameters for comparing varieties. Crop Sci 6:36–40

Eskridge KM (1990a) Selection of stable cultivars using a safety-first rule. Crop Sci 30:369–374

Eskridge KM (1990b) Safety-first models useful for selecting stable cultivars. In: Kang MS (ed) Genotype-by-environment interaction and plant breeding. Louisiana State University, Baton Rouge, La.

Eskridge KM (1991) Screening cultivars for yield stability to limit the probability of disaster. Maydica 36:275–282

Eskridge KM, Byrne PF, Crossa J (1991) Selecting stable cultivars by minimizing the probability of disaster. Field Crops Res 27:169–181

Eskridge KM, Johnson BE (1991) Expected utility maximization and selection of stable plant cultivars. Theo Appl Genet 81:825–832

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

International Maize and Wheat Improvement Center (CIMMYT) (1990) CIMMYT international maize testing program 1988 report. CIMMYT, Mexico, D.F.

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

Lin CS, Binns MR, Lefkovitch LP (1986) Stability analysis: where do we stand? Crop Sci 26:894–899

Menz KM (1980) A comparative analysis of wheat adaptation across international environments using stochastic dominance and pattern analysis. Field Crop Res 3:33–41

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

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

Steel RGD, Torrie JH (1980) Principles and procedures of statistics, 2nd edn. McGraw-Hill, New York

Winer BJ (1971) Statistical principles in experimental design, 2nd edn. McGraw-Hill, New York

Appendix

Testing equality of reliabilities for k test cultivars based on normally distributed differences and the Wald test

Assume test trial data are available for the check and k test cultivars in n environments. Let $d_{ij} = Y_{ij} - Y_{cj}$ be the difference between the response of the ith test cultivar and that of the check in the jth environment. Let the 1 x k vector of differences of the k test cultivars for the jth environment $\mathbf{d}_j = (d_{1j}\ d_{2j} \ldots d_{kj})$ be one of n independent samples from a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mu_{d1}\ \mu_{d2} \ldots \mu_{dk})$ and a $k \times k$ covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ has σ^2_{di} as the ith diagonal element and σ_{dij} as the i, jth off-diagonal element. The marginal distribution for the ith cultivar then has a univariate normal distribution with mean μ_{di} and variance σ^2_{di} (Anderson 1958). Further, define $p_i = P(d_{ij} > 0)$ and note $p_i = 1 - \Phi(-\mu_{di}/\sigma_{di})$ where $\Phi(.)$ is the standard normal distribution function. Then the hypothesis of interest is $H_0: p_1 = p_2 = \ldots = p_k$.

To test H_0 using Wald's test (Nelson 1982), define $k-1$ constraint functions:

$$h_i(\mu_{d1}, \ldots, \mu_{dk}, \sigma^2_{d1}, \ldots, \sigma^2_{dk}) = p_1 - p_i = 0; \quad i = 2, \ldots, k.$$

Obtain a $2k \times (k-1)$ matrix of partial derivatives: $\mathbf{H}(\boldsymbol{\Theta}) = \{\delta h_i/\delta \Theta_j\}$; $i = 2, \ldots, k$ and $j = 1, \ldots, 2k$ where Θ_j is the appropriate parameter μ_{dl}, or σ^2_{dl}. The $2k \times 2k$ covariance matrix of the maximum likelihood estimators ($\hat{\boldsymbol{\Theta}} = (\hat{\mu}_{d1}, \ldots, \hat{\mu}_{dk}, \hat{\sigma}^2_{d1}, \ldots, \hat{\sigma}^2_{d2})$) is defined as $\boldsymbol{\Sigma}(\boldsymbol{\Theta})$ with elements

$$\begin{aligned}
\operatorname{cov}(\hat{\mu}_{di}, \hat{\mu}_{dj}) &= \sigma^2_{di}/n && \text{for} \quad i=j \\
&= \sigma_{dij}/n && \text{for} \quad i \neq j \\
\operatorname{cov}(\hat{\sigma}^2_{di}, \hat{\sigma}^2_{dj}) &= 2(n-1)\,\sigma^4_{di}/n && \text{for} \quad i=j \\
&= 2(n-1)\,\sigma^2_{dij}/n && \text{for} \quad i \neq j
\end{aligned}$$

and all other elements zero.

The asymptotic covariance matrix of the maximum likelihood estimates evaluated at $\hat{\boldsymbol{\Theta}}$ is $\boldsymbol{V} = \boldsymbol{H}(\hat{\boldsymbol{\Theta}})' \boldsymbol{\Sigma}(\hat{\boldsymbol{\Theta}}) \boldsymbol{H}(\hat{\boldsymbol{\Theta}})$. Now define $\mathbf{h}(\hat{\boldsymbol{\Theta}})$ as the $k \times 1$ vector of constraints evaluated at $\hat{\boldsymbol{\Theta}}$, the maximum likelihood estimates. Then Wald's statistic for testing H_0 is

$$W = \mathbf{h}(\hat{\boldsymbol{\Theta}})' \mathbf{V}^{-1} \mathbf{h}(\hat{\boldsymbol{\Theta}})$$

which has a chi-square distribution with $k-1$ degrees of freedom when H_0 is true and the sample size is large (Nelson 1982).

Relationship between stability parameters and variance of the test-check differences

Define the performances of the ith test cultivar and the check in environment j as Y_{ij} as Y_{cj}, respectively.

Following Shukla (1972), for the hth cultivar

$$Y_{hj} = \mu + G_h + E_j + V_{hj} \quad \text{for} \quad h = i \text{ or } c$$

where μ is the grand mean, G_h is a fixed cultivar effect, E_j is an environmental effect and V_{hj} is the h, jth random deviation from the additive model. V_{hj} has expectation 0, variance σ^2_h and covariance $\operatorname{Cov}(V_{ij}, V_{cj}) = 0$. The difference between the ith cultivar and the check environment j is

$$\begin{aligned}
d_{ij} &= Y_{ij} - Y_{cj} \\
&= G_i - G_c + V_{ij} - V_{cj}.
\end{aligned}$$

Using rules of expectation, the variance of d_{ij} over environments is $\sigma^2_{di} = \sigma^2_i + \sigma^2_c$.

For Eberhart and Russell's model (1966), define

$$Y_{hj} = \mu_h + \beta_h I_j + \delta_{hj} \quad \text{for} \quad h = i \text{ or } c$$

where μ_h is a fixed cultivar effect, β_h is the regression coefficient for the hth cultivar, I_j is a random environmental index, and δ_{hj} is the deviation from regression of the hth cultivar in the jth environment. I_j has expectation 0 and variance σ^2_I while δ_{hj} has expectation 0, variance $\sigma^2_{\delta h}$ and covariance $\operatorname{Cov}(\delta_{ij}, \delta_{cj}) = \operatorname{Cov}(I_j, \delta_{hj}) = 0$. Given this model, the difference between the ith cultivar and the check is

$$d_{ij} = \mu_i - \mu_c + (\beta_i - \beta_c)\, I_j + \delta_{ij} - \delta_{cj}.$$

Using rules of expectation, the variance of d_{ij} over environments is $\sigma^2_{di} = (\beta_i - \beta_c)^2\, \sigma^2_I + \sigma^2_{\delta i} + \sigma^2_{\delta c}$.

The relationship between σ^2_{di} and Finlay and Wilkinson's slope parameter (1963) is obtained using the Eberhart and Russell model with all δ's and their variances set to zero.