HILGARDIA

A JOURNAL OF AGRICULTURAL SCIENCE PUBLISHED BY
THE CALIFORNIA AGRICULTURAL EXPERIMENT STATION

Volume 35, Number 22 • November, 1964

Uniformity Field Trials and Monte Carlo Simulations

George A. Baker and John P. Johnson

Significant Differences on the Basis of Stable Rankings Analyzed by the SD Technique

George A. Baker and Burton J. Hoyle

THIS ENDS VOLUME 35

UNIVERSITY OF CALIFORNIA DIVISION OF AGRICULTURAL SCIENCES

In the first paper, actual uniformity field trials are examined and it is found that analyses based on conventional mathematical models may assess very poorly the probabilities used in detecting significantly different varieties.

Monte Carlo results show changes in the mathematical model of field trials that can give probability distributions that correspond closely to the distributions observed for actual trials.

In the second paper, emphasis is placed on reproducibility of field plot results as the most desirable evaluation. Techniques by which a stable ranking among treatments can be obtained (i.e.: A is better than B) are discussed as a matter of field plot manipulation. Examples are given where reproducibility, as measured by the SD technique in a single year, is applicable to a high degree of certainty to results based on several years' experience. The SD technique provides a confidence limit depending on design, and the values of the limits are computed.

A reproducible ranking order is held to be desirable and the problems of securing one are discussed. Techniques are offered which simplify obtaining a stable ranking. Mathematical formulas are given by which given cut-off points of confidence can be calculated. Adequate field plot decisions are based on both agronomic usefulness and mathematical confidence. The SD technique is shown to fulfill both of these considerations.

THE AUTHORS:

George A. Baker is Professor of Mathematics and Statistician in the Experiment Station, Davis; John P. Johnson was at the time of these studies graduate student at the University of California, Davis, and is at present graduate student in statistics at Iowa State University, Ames; Burton J. Hoyle is Specialist in Field Station Administration and Superintendent of the Tulelake Field Station, Tulelake.

George A. Baker and
Burton J. Hoyle

Significant Differences on the Basis of Stable Rankings Analyzed by the SD Technique[1]

INTRODUCTION

WHEN THE SAME group of barley varieties is grown in a nursery a number of times, the varieties may rank differently each time. However, the better varieties are usually found near the top of the list and the poorer varieties in a lower position. Widely different varieties, therefore, are quite stable with respect to each other, but a problem arises when comparing two varieties of similar characteristics, because they so frequently alternate in rank. Assuming that a 50:50 alternation of rank between two similar varieties over a large number of trials indicates equality between the two, then a 100 per cent bias in favor of one over the other indicates strong significance difference between the two. Between these extremes lies an area where a sufficient increase in the number of rankings of one over the other can only mean that one is significantly better. It should be possible, then, to select a succession of varieties, each significantly higher in the sequence, which have a very high stability in ranking order each time they are grown together in a nursery. This means that for each included variety a dominant characteristic such as "total yield" overrides most of the minor uncontrolled variations in moisture, fertility, and others. Conversely, unless a group of varieties can be shown to have a stable ranking over a large number of trials, they cannot be said to be significantly different. A "better treatment" under field conditions, then, might well be one with the characteristic of resisting relative change in ranking due to minor and unknown variations. In this paper, we shall use the term variety as a general term for variety or treatment.

There are many apparent difficulties in obtaining such significant rankings for groups of varieties. In spite of these apparent difficulties, stability of ranking has, in fact, been demonstrated for a group of Hannchen barley mutants (Hoyle and Baker, 1961),[2] also with a group of potato treatments testing seed size and location (Hoyle and Baker, 1960), as well as on lettuce variety and quality trials (Hoyle, 1959). Other developments in the theory and practice of testing by means of field trials are indicated in the listed references.

This study concerns itself, first, with new data demonstrating various aspects of the use of the ranking technique and

The term SD is an abbreviation for the expression "stability of ranking in descending order." For an application of the SD technique see tables 1 and 2.

[1] Submitted for publication April 15, 1964.

[2] Publications named in the text refer to those listed in "Literature Cited" on page 645. Other pertinent publications are listed in "References" on pages 645–46.

with developing a background for establishing significance limits. Second, significance limits are obtained empirically for SD scores defined in a previous publication (Hoyle and Baker, 1961).

MATERIALS AND METHODS

In the 1960 crop season, many six-row barley varieties were grown in various combinations in many nurseries at the Tulelake Station. A grand average for each variety was computed and these varieties ranked accordingly. Three groups of seven each were chosen and the varieties Traill and G×T were added to each. Each nursery was put together so as to form a succession of varieties, each higher in yield by about 10 per cent, as judged from 1960 data, than the next lowest. Some duplication was necessary, as will be shown. During the years 1961 and 1962, each nursery was grown as single 4-foot rows replicated in a 9×9 latin square.

FACTORS AFFECTING RELIABLE RANKING

Obtaining a stable rank among a series of varieties requires more than a superficial look. Chief among the considerations are:

1. Several varieties within a group may be nearly equal and, if so, will form a subgroup. These varieties will rank at random within the limits of this subgroup and can often be identified by a SD score. One goal of analyzing by ranking is to identify such subgroups. Those varieties in such a group are judged to be equal to each other, all having nearly the same numerical value of SD score and all being interchangeable in ranking order.
2. "Maverick" varieties do exist and can be identified by the SD score. These are varieties so unstable, caused by lack of a consistent response to minor changes in environment, that they seldom achieve stability in ranking and serve primarily to cover up an orderly ranking among other treatments. They can usually be identified and removed from the array and then it can be observed whether or not the remaining treatments stabilize in rank. The "maverick" tendency may be genetically or error oriented. The "maverick" behavior may indicate a strong sensitivity to slight changes in environmental conditions and may lead to highly desirable results if properly understood.
3. Lack of over-all stable ranking is most often an indication of a "poor test," that is, one full of variations of all kinds. Such a test will not show significance by a conventional analysis of variance or by a ranking comparison.
4. Ranking provides a system of forming groups of like treatments. Two groups, each significant by a SD score, may be combined to form a larger group of greater significance, but a large group of minimum significance may not be divided into smaller groups.

INTERPRETATION BY THE SD TECHNIQUE

Interpretation of the field plot information in this study was based on the methods outlined previously (Hoyle and Baker, 1961). In that study, three methods of analysis were compared—the Analysis of Variance, Game Theory and the SD Technique. The SD Technique was so promising that further work seemed desirable. The important features of this system, along with an example, are presented in this section. Emphasis is placed on the empirical results of our trials that 'islands' are an inescapable phenomenon of field trials applicable to *all* size plots. For this reason there is no such thing as a uniform plot and we have never observed predetermined blocks, or other designs to include exclusive contiguous levels of productivity—only islands do this, and on a posterior basis. A 'uniform' area with respect to one characteristic, whether an island or a block (if possible), is seldom uniform with respect to all attributes. An area may be uniform for total yield, but nonuniform for protein, for instance. For this reason we consider the term 'uniform' as applied to field plots paradoxical. Nevertheless uniform areas specific for each attribute can be identified by the SD Technique, and are useful as an aid in evaluation.

Rigid block effects are ignored and replaced by the concept that variation occurs in unpredictable islands of variation, crossing any predetermined rows and columns, and suited for approximate determination *after* experimentation. Each of these islands are areas of similar productivity level, and only those plots falling within a given island can justifiably be compared with each other. Varieties are compared at the levels of highest productivity, second highest, and so on.

Standard techniques of field design are used to specify field locations of treatments and the labor required for the eight to twelve locations found desirable is offset by using smaller plots. Upon obtaining a set of data, the values for each treatment are ranked in separate columns. The top value for each variety comes from the island of greatest productivity and is called island 1. The variety with the highest value in island 1 ranks first. The remaining varieties are ranked within island 1 and then the process is repeated for each of the other islands. The best variety then will be found to rank first in nearly all of the islands, or, we might say, it will be found to rank first at nearly each level of productivity. Any two varieties are compared on the basis of rank—regardless of magnitude, and if six field locations with one design have been used, there will be six islands and six areas of comparison.

The SD score is so constructed that its meaning indicates significance on the one hand and provides a means of calculating the degree and amount of deviation encountered in each island from the mean.

CALCULATING THE SD SCORE

The SD score is calculated for a 9×9 latin square experiment as follows: the variety values are ranked in columns as shown in table 1. Variety names are listed as numbers 1 through 9. Table 2 is next constructed, listing the variety names as they ranked in each island of table 1. The rank and the value of the mean for each variety is also listed as shown in table 2. Finally, the variety

TABLE 1

AN EXAMPLE OF VARIETY VALUES ARRANGED IN DESCENDING ORDER

Island*	Treatments								
	1	2	3	4	5	6	7	8	9
1	62	91	38	69	109	71	90	49	106
2	61	89	33	55	99	70	88	42	105
3	60	86	30	49	97	67	87	41	102
4	59	85	29	47	95	65	86	41	102
5	57	84	22	45	91	58	83	41	96
6	54	83	20	44	89	53	82	33	89
7	49	74	18	44	81	51	81	32	82
8	48	71	17	32	80	45	66	22	78
9	38	68	12	28	79	44	63	20	75
Average	54.2	81.2	24.3	45.9	91.1	58.2	80.7	35.7	92.8

* The area of highest productivity for each variety is called island 1, the next island 2, and so on.
For the area of highest productivity, island 1, the varieties rank 5, 9, 2, 7, 6, 4, 1, 8 and 3, which is shown as the first column of table 2.

TABLE 2

STABILITY OF THE VARIETIES OF TABLE 1 WHEN RANKED IN DESCENDING ORDER FROM EACH ISLAND

Rank	Island number									Mean variety		SD score	
	1	2	3	4	5	6	7	8	9	Name	Value	Single	Group*
1........	5	9	9	9	9	9	9	5	5	9	92.8	3	0
2........	9	5	5	5	5	5	7	9	9	5	91.1	4	1
3........	2	2	7	7	2	2	5	2	2	2	81.2	3	0
4........	7	7	2	2	7	7	2	7	7	7	80.7	12	1
5........	6	6	6	6	6	1	6	1	6	6	58.2	2	0
6........	4	1	1	1	1	6	1	6	1	1	54.2	3	1
7........	1	4	4	4	4	4	4	4	4	4	45.9	1	1
8........	8	8	8	8	8	8	8	8	8	8	35.7	0	0
9........	3	3	3	3	3	3	3	3	3	3	24.3	0	0

* Horizontal lines in this column separate varieties into groups. Groups are determined by examination of the data. The LSD as calculated from the analysis of variance is 12.5 for the 1 per cent level.

SD scoring method:

Each mean variety name in agreement with an island rank scores 0.

Each deviation by one rank scores 1, each deviation by over one rank scores 10. The group score is obtained the same way except the agreement is for within the subgroup, and not for a single rank.

name in the mean column is compared with the rank position it occupies in each island, the score is a measure of how it deviates, and the scoring procedure is indicated at the bottom of table 2. Establishing significance limits for the SD score will occupy the final section of this paper.

RESULTS AND DISCUSSION

ABSOLUTE RANKING ORDER

The high degree of success in obtaining a stable ranking over a three-year period is noted in table 3. Of the 27 varieties drawn from three nurseries, six ranked the same each year, nine deviated by one rank only in one of the years, nine deviated by two ranks in at least one of the years, and the final three deviated by three ranks only. These initial groups were picked according to an estimated scale and therefore do not represent a random choice of varieties. However, the 1960 list was made up of random varieties submitted from various sources to the station. The years 1961 and 1962, then, have confirmed strengths and weaknesses of certain selections we were interested in. The following is of interest in table 3:

1. In BR1, varieties 2 and 3 were initially nearly the same in yield, as were varieties 4 and 5. Their rank over the years should alternate according to theory if this near equality is a fact. Such was the case in this limited example. Variety 6 is a "maverick" and, by its removal from the summary of ranks, the over-all agreement in rank is improved somewhat.

2. In BR2, varieties 4 and 5 were nearly the same in 1960. By dropping out 5 as a duplicate in value, the overall ranking is improved. Stable ranking depends on lack of nearly equal values among the varieties.

The above comments raise the question, how large must variety differences be in order to achieve stable ranking? This will presumably vary with location and many other factors, but the varietal differences observed in the three nurseries increased as follows, from the lowest to the highest value:

TABLE 3

AVERAGE YIELD AND RANK OF BARLEY VARIETIES GROWN OVER THREE YEARS. VARIETIES WERE SELECTED FOR STABILITY IN 1960

Varieties	1960		1961		1962		Summary of ranks		
	Yield	Rank	Yield	Rank	Yield	Rank	1960	1961	1962
Nursery BR1									
Traill	251	1	272	1	257	1	1	1	1
4706	206	2	199	3	222	3	2	3	3
Hannchen	197	3	146	6	194	5	3	6	5
G × T	179	4	268	2	229	2	4	2	2
5348	177	5	190	4	214	4	5	4	4
Atlas 46	156	6	187	5	170	7	6	5	7
855	143	7	135	7	91	9	7	7	9
Parkland	124	8	111	8	179	6	8	8	6
Forrest	105	9	62	9	140	8	9	9	8
Nursery BR2									
Traill	251	1	297	1	258	1	1	1	1
3317	215	2	277	2	244	2	2	2	2
4363	196	3	261	3	239	3	3	3	3
G × T	179	4	233	4	221	4	4	4	4
446	179	5	195	6	217	5	5	6	5
1019	162	6	170	7	205	6	6	7	6
298	144	7	203	5	192	7	7	5	7
570	129	8	153	9	146	8	8	9	8
Bonn	121	9	160	8	83	9	9	8	9
Nursery BR3									
Traill	251	1	324	1	201	2	1	1	2
3317	215	2	242	4	210	1	2	4	1
Firlbeck	197	3	263	2	179	4	3	2	4
1163	184	4	225	5	161	6	4	5	6
G × T	179	5	262	3	194	3	5	3	3
Montcalm	167	6	155	7	136	8	6	7	8
Atlas 57	148	7	213	6	139	7	7	6	7
855	143	8	103	8	168	5	8	8	5
Forrest	105	9	70	9	115	9	9	9	9

BR1
18.0, 15.3, 9.0, 13.3, 1.1, 10.0, 4.5, 21.8%
with average 11.6%

BR2
6.6, 11.6, 12.5, 10.4, 0.0, 9.4, 9.6, 16.7%
with average 9.6%

BR3
37.1, 3.4, 12.8, 7.1, 2.7, 7.0, 9.1, 16.7%
with average 12.0%

By disregarding the highest and lowest values of these series as atypical, the average differences between variety values are 8.9, 8.9 and 7.0 per cent respectively. Least stability was found at 7.0 per cent from BR3. The above values did not vary greatly from year to year, being more closely correlated with nurseries. These results leave little doubt that it is possible to easily segregate varieties with differences of less than 10 per cent into highly reproducible ranking patterns. This fact should not be taken lightly, because primarily a reproducible phenomenon is, after all, the best test of a real difference, and many of these reproducible events are indicated as nonsignificant by the analysis of variance.

If stable ranking over a three-year period provides a good measure of confidence with no further proof, then the most serious question to be answered is how to handle 57 or 157 varieties at once instead of nine at a time. One answer is to simply break these large numbers down into small groups of nine or 12 and gain the confidence needed to realize that the highest island is roughly the same for each nursery, and that its absolute value can be adjusted with practice to fit most cases. It may actually become possible to compare varieties directly, even though they are not in the same nursery.

Of most immediate concern is the question of whether or not it is possible to detect strong, significant trends between varieties during the first or second year. This is done in the following manner.

RANKING BY SD GROUPS

By taking the results of table 3 and performing the SD calculations on them, the data can be rearranged as seen in table 4. The rules stipulate that each group has a common numerical value, and the included varieties are interchangeable as to rank and value. The SD score is individual for each variety with reliability decreasing when the score increases. Nonsignificant varieties should be dropped from their groups and considered as nonentries in the test.

Considering table 4 further, there are several points of interest. Observe that each nursery has the same number of varieties but a different number of similar groups. It would be desirable to have number of varieties equal number of groups. Because there is more than one variety in a group, we know they alternated in rank through their various levels of productivity where they were compared. In some instances we have combined significant groups for the

TABLE 4

SD GROUPS FORMED BY YEARS FOR THREE NURSERIES

Group	1960			1961			1962		
	Variety	SD score	Yield value	Variety	SD score	Yield value	Variety	SD score	Yield value
Nursery BR1									
1	Traill	..*	251	Traill	0	270	Traill	11	257
				GxT	0				
2	4706	..	202	4706	0	192	4706	2	226
	Hannchen	..		5348	0		GxT	5	
				Atlas 46	0				
3	GxT	..	178	Hannchen	1	141	Hannchen	2	204
	5348	..		855	0		5348	53	
4	Atlas 46	..	150	......	..	...	Atlas 46	3	175
	855	..		......	..	...	Parkland	1	
5	Parkland	..	124	Parkland	0	111	Forrest	0	140
6	Forrest	..	105	Forrest	0	62	855	0	91
Nursery BR2									
1	Traill	..	221	Traill	0	278	Traill	30	247
	3317	..		3317	2		3317	0	
	4363	..		4363	1		4363	2	
2	GxT	..	173	GxT	2	210	GxT	3	221
	446	..		446	0		446	43	
	1019	..		298	11				
3	298	..	131	1019	1	161	1019	14	199
	570	..		570	0		298	2	
	Bonn	..		Bonn	0		570	0	146
							Bonn	0	83
Nursery BR3									
1	Traill	..	221	Traill	0	283	Traill	1	202
	3317	..		GxT	1		GxT	12	
	Firlbeck	..		Firlbeck	1		3317	0	
2	1163	..	177	1163	0	227	Firlbeck	3	169
	GxT	..		3317	11		855	12	
	Montcalm	..		Atlas 57	0		1163	3	
3	Atlas 57	..	146	Montcalm	0	184	Atlas 57	12	138
	855	..		855	0		Montcalm	4	
4	Forrest	..	105	Forrest	0	70	Forrest	4	105

* No SD score was computed in 1960. Yield values were used in 1960 to form the initial groups.

sake of clarifying the over-all data, as in group 3, 1962 for BR2. The over-all group values here would be 143 for the four included varieties. Bonn was significantly lower at 83, and variety 570 lower at 146. These atypical values give a distorted field value which is questionable but, as noted in table 3, the same low values did not disturb the ranking stability.

From table 3 we know what absolute rank a variety occupied, but there is no way to determine with which other varieties it shared a common value. This information is actually given by the groupings of table 4. A variety's relative associations with other varieties over the years can be observed.

In all cases of table 4, a SD value of 15 or over appears to indicate nonsignificance, as noted in our observations of many sampling results. It has been observed that significance determined on this basis is: (1) in high agreement with a realistic analysis of variance, factor analysis, and game theory; (2) a phenomenon readily reproduced in repeated field trials; (3) much easier and more quickly applied than other methods.

There can be little doubt from practical experience that the SD score expresses conditions where the mean differences so indicated are different beyond the element of chance so far as rank is concerned. This fact is also expressed in theory, and the following section presents the background for this.

BACKGROUND FOR SIGNIFICANCE LIMITS FOR SD SCORES

The SD score reveals more about a field plot than a statement as to whether or not a difference could be due to chance alone. Consider a chalk line marked on a table on which coins are being tossed. The line represents a level of performance. The SD score can best be described as: (1) giving us a measure of the *frequency* of heads vs. tails where heads or tails represent kinds of genetic or physiological plants; (2) telling us *how close* the coins fall to the chalk line in terms of near or far; (3) *how many* of the coins fall near and how many fall far.

For this discussion, the 9×9 latin square is used as an example for which significance limits are established. Significance limits for this square are established by two approaches. The first is for the restricted case of a 9×9 square, and the second is a more general and theoretical viewpoint applicable to any rectangular arrayed experiment. The latin squares are discussed because of the rather considerable background material available to work with in both areas of field trials and theory. It is not to be implied that the SD score is restricted to squares. There is much evidence that the scoring and theory applies equally well to many other good field designs.

STRUCTURE OF THE SCORE

Since a variety mean determines or "fixes" which rank it will occupy within the group of varieties (see the mean-variety column of table 2), we can argue that the elements of chance were greatly restricted or absent in this choice. Of course the elements of chance were present in each of the islands of productivity, which indirectly determines the means. Thus, in a most restricted case, we can visualize but two means with their nine respective values, as in the following case, 8 and 3 being variety names:

Rank	Islands									Variety Mean	Variety Name	SD Score
	1	2	3	4	5	6	7	8	9			
1.....	8	8	8	8	8	8	8	8	8	35.6	8	0
2.....	3	3	3	3	3	3	3	3	3	24.5	3	0

The only possible SD scores which could be derived from such a restricted array would be 0, as shown in the above table, or 1, 2, 3, 4, 5, 6, 7, 8. A score of 9 would not be possible with only two varieties since this would completely reverse the order of the varieties, and this cannot be done because the means are fixed. For this restricted example, then, of only two varieties, it is possible to calculate the frequency with which each could occur, as follows:

Score	Frequency	Approximate Significance (%)
0	1	0.2
1	9	1.7
2	36	7.0
3	84	16.3
4	126	24.9
5	126	24.9
6	84	16.3
7	36	7.0
8	9	1.7
Total	511	

From the above table we note that there is only one way out of 511 chances where nine islands would rank as our treatments 8 and 3 do previously. There would be nine different combinations giving a score of 1, 36 combinations providing a score of 2, etc. Viewed with the restrictions as stated, a SD score of 0 provides a significance level of 0.2 per cent, a SD score of 1 provides a significance level of 1.7 per cent, a SD score of 2 has 7.0 per cent significance, etc.

The frequencies do not increase uniformly with the scores. This has *real meaning* in terms of actual field conditions. For example, scores 1 and 8 each have an expected frequency of 1.7 per cent. The meaning, in *addition* to the significance levels, is illustrated in the two cases on the bottom of the page.

In case 2 it is apparent that the lone value of 8 in rank 1 would need to be extremely high in order for the mean value of 8 to exceed that of treatment 3. Equal distortion would need to be in the lone value of treatment 3, rank 2. Such distortions make a field plot immediately suspect and in fact rarely exist. For this reason, scores of 8 (and 7) are rarely found in actual trials, but when they are, they are indeed significant, but in a negative way. In our case 2, below, the score of 8 is indeed significant at the 1.7 per cent level of confidence, but for the ranking of 3 over 8 and not the ranking of 8 over 3, as indicated by the two mean values.

Every SD score is meaningful for a distinctive pattern of field behavior which can be useful to the experimenter, but only scores of 0, 1 and 2 reflect confidence at the usually accepted levels and for the limited case just described. Where only two lines of ranks are compared, all disagreements are by one rank only and are considered as near misses. The usual situation will contain opportunity for far misses in addition, and significance under these conditions will now be discussed.

Case 1: Usual

Rank	Islands									Mean	SD Score	
	1	2	3	4	5	6	7	8	9			
1	8	8	8	8	8	8	8	8	3	8	1	High Mean Value
2	3	3	3	3	3	3	3	3	8	3	1	Low Mean Value

Case 2: Unusual

1	3	3	3	3	3	3	3	3	8	8	8	High Mean Value
2	8	8	8	8	8	8	8	8	3	3	8	Low Mean Value

COMPUTING SIGNIFICANT DIFFERENCES FOR SD SCORES

The point of view taken for evaluating field plots by the SD technique is that a predetermined matrix is constructed in such a manner that when field data are entered in the way described, the score determined is judged with respect to a precalculated level of significance. In all but 2×2 squares, deviations from a mean position are measured both by one and by more than one rank. The larger the square, the greater total number of scores possible and the greater the complexity in computing the significant level of each. Squares are by no means necessary for computing SD scores, but are used in this study because of the large amount of background material available. There are several approaches for determining the significance of each SD score, depending on the assumptions and methods employed. Each method has points of similarity which recur in each approach. This section explores two approaches.

1. A RESTRICTED 9×9 SQUARE AS IN TABLE 2—THE THREE-LINE CONCEPT

In a 9×9 square there are 55 possible SD scores, as shown in the array of table 5*a*. By imposing restrictions based on two assumptions, an example of the frequency distribution of each score is given. The two assumptions are:

A. That any three lines are a unit, as, for instance,

Lines	Islands								
	1	2	3	4	5	6	7	8	9
1........	5	5	5	5	5	5	5	5	5
2........	3	3	3	3	3	3	3	3	3
3........	8	8	8	8	8	8	8	8	8

In this version, both far and near deviations are possible, and all scores which are possible in a 9×9 square are also possible in this grouping. The three lines in this concept are symbolic in that the third line is representative of all far deviations, regardless of how far away.

B. Also assumed is that *one* of the directional deviations, i.e., from top to bottom or bottom to top, but not both, is used. In this array as in all SD scoring, each line (or group) is evaluated as a single entity, and the treatment is seen as a group of points overlapping contiguous groups but most stable in the position occupied. In the above grouping there are 27 points displayed while in the full 9×9 square there are 81 points. Therefore, the assumptions set forth ignore the possibility that increasing the number of lines also increases the number of points which need to be considered for determining significance.

There is some validity in making the above assumptions. First of all, the SD score is based on near misses which are always by one rank only, and on far misses based on more than one rank displacement. Thus, all far misses are equal, regardless of magnitude. In the hierarchy of rank, the magnitude of the mean values fixes their location in the order of ranking through mechanical and not random techniques. On the other hand, the magnitude of the mean is directly determined by random field values. Where field values for varieties represent corresponding levels of productivity, the individual values are governed, or affected by, the varieties which reflect a distinct clustering pattern of values, different for each variety. In other words, there is a natural restriction of random field values imposed by a variety. This restriction has a tendency to prevent random placement throughout the whole matrix. Thus, for certain groups of varieties, a significance based on scoring by the three-line concept may be completely valid.

Consider an extreme test containing 30 varieties with a four- or fivefold range of output values. The top three varieties would seldom occur in the bottom 3 ranks because of their nature. Just as three jet airplanes would almost always win a race with three biplanes because of their nature. In the air race there would be six ranks to fill of the winning order and yet we can hardly say that a single biplane has a 1-in-6 chance of winning just because it is in the race. If our thirty varieties are vegetables and as mismatched as the air planes we cannot say that the number of possibilities of winning is truly 1 in 30 just because a particular vegetable is in the trial. This is what we mean when we say that there is a natural restriction imposed by a variety, and any group of 2, 3, or 4 may best be evaluated as a subgroup of the thirty. This element of natural restriction removes the possibility of complete randomness in any ranking order and for this reason partially justifies the use of confidence limits based on a restricted number of lines rather than being forced to use all lines involved.

Here is an example:

Relative frequencies of the restricted three-line concept

In table 5*a* the triangular array displays the 55 possible SD scores when at least 3 treatments and 9 replications are used. Each line contains only those scores having the same number of total deviations of all kinds. The sixth line shows scores of 5, 14, 23, and so on. If the digits of each score are added together, they total the same for all scores on a line, i.e., $5+0=5$, 14 is $1+4=5$, 23 is $2+3=5$, etc.

TABLE 5*a*

TOTAL NUMBER OF SD SCORES POSSIBLE IN ALL MATRICES FROM 3×9 TO 9×9

0									
1	10								
2	11	20							
3	12	21	30						
4*	13	22	31	40					
5	14	23	32	41	50				
6	15	24	33	42	51	60			
7	16	25	34	43	52	61	70		
8	17	26	35	44	53	62	71	80	
9	18	27	36	45	54	63	72	81	90

Each score in a given line represents the same total number of deviations but describes a different agronomic pattern.

* Each column is separated by a line into an upper and lower half.

In table 5*b* the number of ways each score can occur is shown in column 1, and scores with similar frequencies are grouped on each line. There are a total of 19,683 different arrangements possible by which all scores can occur. The score of 33 alone can be formed in 1,680 different ways, while the scores of 0, 9 and 90 can be formed in one way only for each, out of the 19,683 total.

Significance of the SD scores using the three-line-concept

The second column of table 5*b* was computed by multiplying the frequency of column 1 times the number of scores shown in the final column and dividing by 19,683 times 100. Thus, column 2 indicates the probability each score in that line has of occurring by chance alone. Line 6, which is underlined, indicates a cut-off point of 3.89 per cent significance, and all scores above this line are significant at the levels indicated. Had we chosen a cut-off point from column 1 only, the dotted line would indicate a 5 per cent cut-off at a frequency of 271 (5 per cent of the total of column 1). These two methods agree very closely.

From table 5*b* it will be noted that both low and high scores are highly significant. As the score goes lower, the

TABLE 5b

COMPLETE SD SCORE DISTRIBUTION BASED ON THE THREE-LINE CONCEPT

Frequency*	Per cent by chance†	0 to 4	5 to 9	10 to 14	15 to 18	20 to 23	24 to 27	30 to 33	34 to 36	40 to 42	43 to 45	50 to 52	53 to 54	60 to 61	62 to 63	70 to 71	72	80	81	90	n
1	.015	0	9																	90	3
9	.26	1	8	10	18													80	81		6
36	1.11	2	7			20	27									70	72				6
72	1.11			11	17											71					3
84	2.59	3	6					30	36					60	63						6
126	3.89	4	5							40	45	50	54								6
252	7.78			12	16	21	26							61	62						6
504	15.56			13	15			31	35			51	53								6
630	9.73			14						41	44										3
756	11.67					22	25					52									3
1260	38.91					23	24	32	34	42	43										6
1680	8.65							33													1
5410																					

* Total arrangements possible are 19,683, or (frequency of line 1 x n) + (frequency of line 2 x n) . . . + (frequency of line 12 x n).

† $\frac{n \times \text{frequency}}{19{,}683} \times 100.$

NOTE: Discussion of the various divided areas is under the section "Significance of the SD scores using the three-line concept" on pages 636 and 638.

If the stair-stepped area is considered as containing the only agronomically acceptable scores, then there are 785 possible arrangements enclosed, out of the total of 19,683. This amounts to 3.98 per cent.

more significant are *differences* between varieties. Conversely, as the score goes higher, the more significant is the fact that there are *no differences* between varieties. It will be noted in table 5*a* that each group of scores, 0 to 9, 10 to 18, 20 to 27, etc., are separated into the top and bottom half. The top half of each group contains those scores where less than half the replicated field values deviate from the mean rank. The second half of each group applies where over half of each group deviates from the mean. It is suggested here that significant variety differences be interpreted as those conditions where the score is less than 5 per cent, table 5*b*, column 2, and also that the scores of more than one digit do not total more than 4. This decision is based on observation of many trials and satisfies primarily an agronomic acceptability. The area mentioned is outlined in table 5*b*. The values 12, 16 and 21 are borderline cases and can be accepted at the significance of 7.78 per cent, if desired. The score 26 signifies a total of 8 out of 9 deviations possible, and is questionable from an agronomic standpoint. Scores above 40 are completely unacceptable for agronomic reasons unless one chooses to use the scores to indicate degree of homogeneity of varieties. Most scores in the 20's and 30's are unacceptable because their probability of occurrence is so high. It is interesting that a diagonal line from the lower left to the upper right of table 5*b* very nearly separates the significant differences from all other scores. Also, the scores of 18, 27 and 36 which all signify total disagreement from the mean rank can be separated by the line as shown.

Figure 1 provides additional insight into the character of the SD score. When plotted as shown, the pathway of consecutive scores forms a nonlinear progression up to about 45, when the pattern changes towards a straight line. The heavy triangle outline marks the area of significance as laid down earlier in table 5*b*.

SD scores describe various structures of agronomic behavior, in addition to significance. Scores of 3, 12, 21 and 30 signify 4 different kinds of field variations. From an agronomic standpoint, 3 is more desirable than 30, and the others are intermediate. In certain instances, scores of 9 and 90 could exist only theoretically, since they indicate a total displacement of field-plot replications from their mean ranking position. For similar reasons, scores of 18, 27, 36, 45, 54, 63, 72 and 81 rarely exist and would fit very peculiar agronomic patterns.

Attention is directed once again to table 4, where scores of above 11 are very rare in actual field plots.

2. MONTE CARLO APPROACH

In this section we obtain empirical sampling distribution of SD scores for 3×3, 6×6, 9×9 and 12×12 latin squares. We use actual uniformity data which is sometimes rather sketchy and hence partially duplicated in some of the squares to find corresponding distributions of SD scores for actual field data under the null hypothesis. We find simple formulas for the means, variance, and 5 per cent significance points in terms of the size of the squares that can be used for interpolation and mild extrapolation. We also compare variability between sets of sample values to variability within sets of sample values.

In the previous section we imposed restrictions which we felt had some justification. In this section, no restrictions are imposed and complete freedom of filling in a matrix is allowed by the Monte Carlo method. That is, in the 9×9 square, neither two-line nor three-line restrictions are imposed, and scoring is permitted in both up and down directions. This has the effect of removing the two-tailed effect noted in table 5*b* where the lowest and highest values have corresponding levels of significance. In interpreting field experiments, the results of the present section can be applied to determine whether further

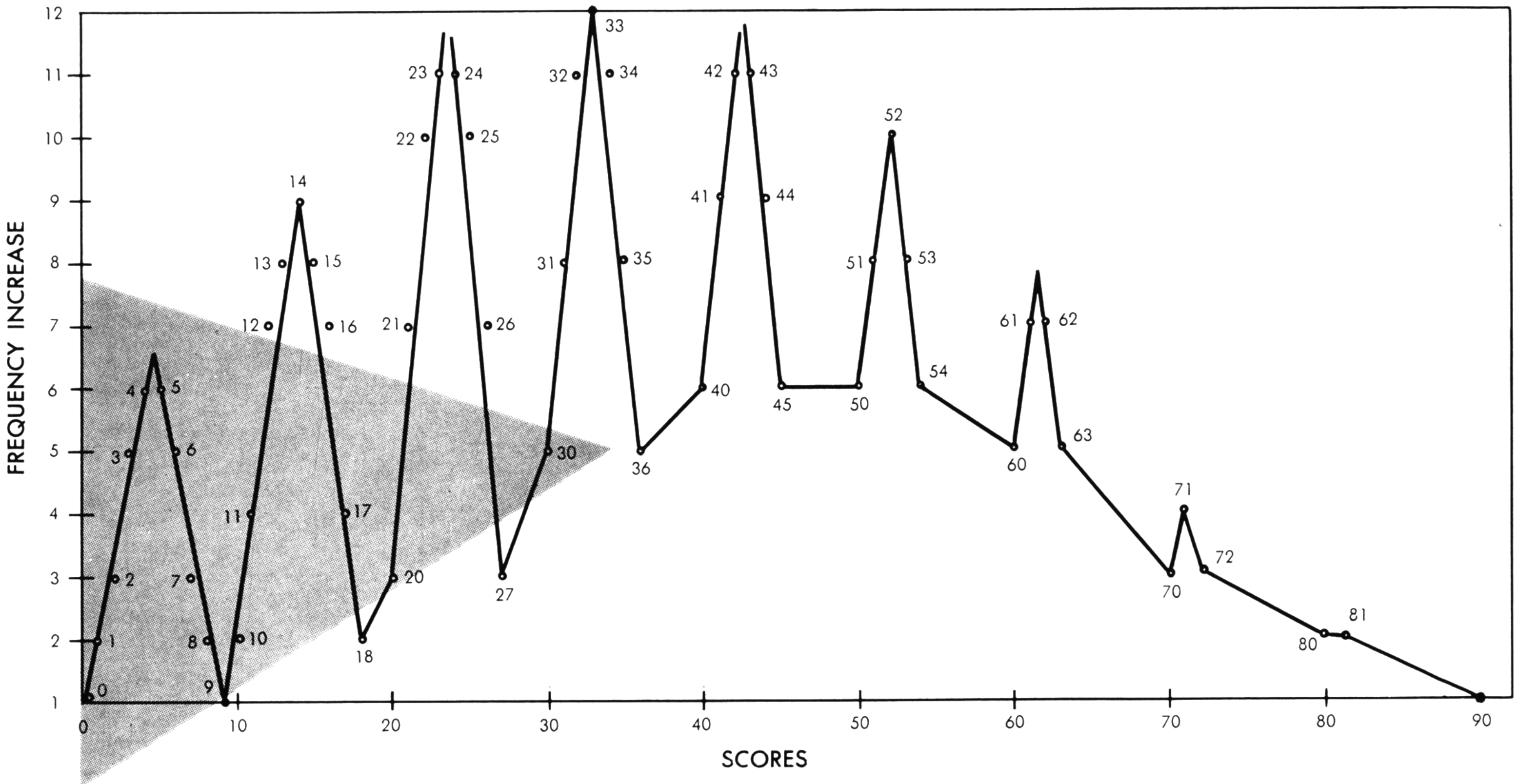

Fig. 1. Showing the consecutive pathway of all possible SD scores in a 9×9 square as related to their frequency of occurrence. See table 5*b* for frequency values.

detailed examination is warranted. If significant differences are indicated on a gross, over-all basis, then further consideration based on the principles explained in the previous sections becomes desirable to explain more fully the actual agronomic situation.

Reference to table 6*c* shows that the theoretical distributions of possible SD scores can be grouped as follows:

Scores	0–29	30's	40's	50's	60's	70's	80's	90's
Distribution	16	55	152	307	401	342	157	28

This shows that the probability of the scores increases as the scores increase through the 60's and then tapers off. Reference to table 5*b* shows that this is the same general pattern for the three-line concept except the change comes in the 50's instead. We also noted earlier in the two-line example this same general pattern. Another point of similarity is noted as follows. Let the arrows denote the direction of increase in probability. Then there is an increase in frequencies from 30 through 33, 34 is roughly similar to 33, and then there is a decrease in frequencies through 36.

30		40	45	50		60	63	70	↑72
31	36	41	44	51	54	↓61	↓62		73
32	35	↓42	↓43	↓52	↓53				
↓33	↓34								

This pattern of theoretical scores corresponds very closely to that of figure 1, which depicts the pathway for the three-line concept. This same pattern is present among the uniformity trial data of table 6*c*.

In calculating the frequency of scores in the full, unrestricted matrix of a square, the higher scores, especially above 27, have a tremendous number of different ways in which they can be formed. Scores below 9 have a built-in restriction limiting them to a one-rank deviation, regardless of how large the square. By imposing a 5 per cent cut-off point at score 34 of table 6*c*, it would appear that this significance level was somewhat higher than that of score 11 of the three-line concept. However, for agronomic reasons a score is not acceptable unless, first, it is within the 5 per cent cut-off area, and second, the numbers of any two-digit partial scores add up to less than 4. This restriction has been determined by numerous field trials. Thus, from the Monte Carlo distribution of table 6*c*, the only acceptable scores would be 1 through 12, 20 and 21 and 30, which is very nearly the same as arrived at by the three-line concept.

Admittedly this approach does not solve all the problems connected with the SD score, but it does point out very definitely that:

1. Very low scores are very highly significant, as determined by several methods.
2. Each approach does have certain elements of similarity.
3. Restricted three, four, or more, concepts may be adequate for determining significant differences.
4. The high scores can be useful as measure of homogeneity.

FREQUENCY DISTRIBUTION

The frequency distributions for Monte Carlo squares and the uniformity trial

TABLE 6*a*

MONTE CARLO AND UNIFORMITY TRIAL DISTRIBUTIONS OF SD SCORES FOR 3 × 3 LATIN SQUARES

Possible SD scores	Monte Carlo distribution	Theoretical distribution*	Uniformity trial distribution
0	59	51.2	6
1	158	155.2	13
2	76	77.2	10
3	10	3.7	6
10	68	77.6	4
11	58	67.9	9
12	7	3.7	..
20	23	22.2	..
21	3	3.7	..
30	..	..	..
Number of samples	154	..	16
Number of scores	462	462	48
Mean	4.90	5.20	3.96
Variance	32.09	32.25	17.96

* Computed by L. J. Brown.

TABLE 6*b*
MONTE CARLO AND UNIFORMITY TRIAL DISTRIBUTIONS OF SD SCORES FOR 6 × 6 LATIN SQUARES

Possible SD scores	Monte Carlo distribution	Uniformity trial distribution
0	..	1
1	..	1
2	6	2
3	9	2
4	6	3
5	3	1
6	1	..
10	3	1
11	13	4
12	27	6
13	31	4
14	9	4
15	3	2
20	10	4
21	66	4
22	87	12
23	51	8
24	18	2
30	32	2
31	98	5
32	96	6
33	35	3
40	60	1
41	101	2
42	32	2
50	54	1
51	43	1
60	12	
Number of samples	151	14
Number of scores	906	84
Mean	30.89	21.01
Variance	147.81	122.15

squares are given in tables 6*a*, 6*b*, 6*c* and 6*d*. Random sampling numbers were used to fill in tables of ranks similar to table 2, and the average rank was used in place of the rank of the average for the variety, as indicated for table 2. There were no values to average for the Monte Carlo scores, so that there could not be ranked averages. Ties in rank are not permitted. Such ties rarely occur, but in case of necessity, ties are broken randomly. The uniformity trial data was somewhat sketchy and partially duplicated to increase the number of trials as far as possible.

The mean and variances for the frequency distributions are given, as are the number of squares and SD scores.

It is clear that all values of the SD score are not possible. For instance, for a 3×3 square, the possible values are 0, 1, 2, 3, then a jump to 10, 11, 12 and so on. Scale considerations of determining significance limits for various sized squares or number of islands led to a uniform value of 10 for far away variations.

There is a general correspondence between the distributions of the SD scores for the Monte Carlo squares and the uniformity trial squares.

The analyses of variance for the SD scores between and within squares are given in tables 7*a* and 7*b*. The general impression is that the SD scores for each square can be approximately considered as a random sample from the over-all distributions given in tables 6*a*, 6*b*, 6*c* and 6*d*.

MATHEMATICAL RELATIONSHIPS

For the purposes of interpolation and possible mild extrapolation, it is desirable to have mathematical formulas. We find:

1. The means of the Monte Carlo distributions are given in terms of the size of square (number of items along one side) by the formula
   square root of the mean = 0.834 (size of square)
2. The variances are given by the formula
   variance = 18.793 (size of square)
3. The value of the 5 per cent significance cut-off point is given by the formula
   5 per cent cut-off point = −40.443 + 8.817 (size of square)

These equations are established by usual least-square techniques from the values found for the empirical frequency distributions.

As further mild extrapolation, it is suggested that the above equations can be stated in terms of the number of islands if the number of varieties is large enough.

TABLE 6c

MONTE CARLO AND UNIFORMITY TRIAL DISTRIBUTIONS OF SD SCORES FOR 9 × 9 LATIN SQUARES

Possible SD scores	Monte Carlo distribution	Uniformity trial distribution
0	..	1
1	..	..
2	..	..
3	..	..
4	..	1
5	..	..
6	..	1
7	..	..
8	..	..
9	..	..
10	..	..
11	..	1
12	1	2
13	..	..
14	2	2
15	..	..
16	..	2
17	..	..
18	..	..
20	..	..
21	..	2
22	2	2
23	1	1
24	7	2
25	2	3
26	1	2
27	..	..
30	..	1
31	3	2
32	10	3
33	17	8
34	17	5
35	7	4
36	1	6
40	2	..

TABLE 6c—*Continued*

Possible SD scores	Monte Carlo distribution	Uniformity trial distribution
41	18	5
42	42	7
43	50	12
44	29	6
45	11	2
50	12	2
51	64	5
52	105	9
53	89	12
54	37	3
60	39	1
61	141	4
62	168	6
63	53	6
70	71	..
71	175	3
72	96	4
80	70	1
81	87	5
90	28	..
Number of samples	162	16
Number of scores	1458	144
Mean	60.97	44.42
Variance	178.66	280.46

TABLE 6d

MONTE CARLO AND UNIFORMITY TRIAL DISTRIBUTIONS OF SD SCORES FOR 12 × 12 LATIN SQUARES

Possible SD scores	Monte Carlo distribution	Uniformity trial distribution
0–13	..	..
14	..	1
15	..	..
16	..	1
17–24	..	..
25	..	1
26–32	..	..
33	..	1
34	..	..
35	..	2
36	..	1
41	..	1
42	..	1
43	..	..
44	1	..
45	..	2
46	1	3
47	..	..
48	..	..
50	..	..
51	1	1
52	1	..
53	4	3
54	1	3
55	3	7
56	2	..
57	2	1
60	1	..
61	..	..
62	8	1
63	16	5
64	10	2
65	11	4
66	..	1
70	2	1
71	13	1
72	27	3
73	25	3
74	28	5
75	7	2
80	5	..
81	40	2
82	66	7
83	66	6
84	28	2
90	30	..
91	84	3
92	116	4
93	46	4
100	46	..
101	98	3
102	69	4
110	72	1
111	63	1
120	39	2
Number of samples	86	8
Number of scores	1032	96
Mean	91.57	70.80
Variance	206.91	445.68

TABLE 7a

ANALYSES OF VARIANCE FOR SD SCORES BETWEEN SQUARES AND WITHIN SQUARES FOR MONTE CARLO SQUARES

Score	SS	df	MS	F
	3 × 3			
Total	14,793.62	461	32.09	
Between squares	6,060.95	153	39.61	1.40
Within squares	8,732.67	308	28.35	
	6 × 6			
Total	133,772.29	905	147.81	
Between squares	26,683.96	150	177.89	1.25
Within squares	107,088.33	755	141.84	
	9 × 9			
Total	260,302.79	1457	178.66	
Between squares	29,092.57	161	180.70	1.01
Within squares	231,210.22	1296	178.40	
	12 × 12			
Total	213,326.55	1031	206.91	
Between squares	20,240.96	85	238.13	1.17
Within squares	193,085.59	946	204.10	

TABLE 7b

ANALYSES OF VARIANCE FOR SD SCORES BETWEEN SQUARES AND WITHIN SQUARES FOR UNIFORMITY TRIAL SQUARES

Score	SS	df	MS	F
	3 × 3			
Total	843.92	47	17.96	
Between squares	358.59	15	23.91	1.58
Within squares	485.33	32	15.17	
	6 × 6			
Total	10,138.99	83	122.15	
Between squares	2,093.82	13	161.06	1.40
Within squares	8,045.17	70	114.93	
	9 × 9			
Total	40,106.16	143	280.46	
Between squares	7,923.27	15	528.22	2.10
Within squares	32,182.89	128	251.43	
	12 × 12			
Total	42,339.24	95	445.68	
Between squares	2,854.49	7	407.78	< 1
Within squares	39,484.75	88	448.69	

SUMMARY

Confidence levels of field trials are normally dependent upon conducting an experiment first and then subjecting the results to a rigorous evaluation. In this paper we have presented a system, called the SD Technique, by which a predetermined method of evaluation, applicable to a field-plot situation, uses predetermined confidence levels. Based on a system of ranking, the data easily and quickly can be placed in a standard array, scored, and the confidence level obtained by reference to a chart.

Data from both actual and theoretical field trials are used and the reliability of the predetermined confidence levels is examined from the standpoint of both theoretical and empirical viewpoints. A discussion is included on the several problems involved with ranking, and suggestions given on how to handle these.

We have considered the distributions of SD scores for hypothetical Monte Carlo and uniformity trial data for latin square designs 3×3, 6×6, 9×9, and 12×12, and have shown a general correspondence between the respective distributions. We have shown that the SD scores for a particular square can be regarded as a sample from the overall distribution. Also, we found simple formulas for the means, variances and 5 per cent cut-off points of the Monte Carlo distributions in terms of the size of the latin squares.

ACKNOWLEDGMENTS

The authors wish to extend their thanks to Mrs. Frances R. Jones for making the Monte Carlo calculations and to Mr. Lloyd J. Brown for some of the theoretical developments.

LITERATURE CITED

HOYLE, B. J.
1959. Tulelake field station, 13th annual report—research. Mimeo, 20 pp.
HOYLE, B. J. and G. A. BAKER
1960. Determining the most efficient plot size for potato trials. Amer. Soc. Hort. Sci. Western Meetings, Eugene, Ore., Mimeo.
1961. Stability of variety response to extensive variations of environment and field plot design. Hilgardia 30:365–94E.

REFERENCES

BAKER, G. A.
1941. Fundamental distributions of errors for agricultural field trials. Nat. Math. Mag. 16: 7–19.
1944. F values for samples of four and four from populations which are the sum of two normal populations. Nat. Math. Mag. 19:62–63.
1945. Test of the significance of the differences of per cents of emergence of seedlings in multiple field trials. J. Am. Stat. Assn. 40:93–97.
1952. Uniformity field trials when differences in fertility levels of subplots are not included in experimental error. Ann. of Math. Stat. 23:289–93.
1952. Field trial problems. Ann. of Math. Stat. 23:480. Abstract.
BAKER, G. A. and R. E. BAKER
1953. Strawberry uniformity yield trials. Biometrics 9:412–21.
BAKER, G. A. and F. N BRIGGS
1945. Wheat bunt field trials. J. Amer. Soc. Agron. 37:127–33.
1949. Wheat bunt field trials. II. Proc. of Berkeley symposium on math. stat. and prob., 1945–46, 485–91.
1950. Yield trials with back-cross derived lines of wheat. Ann. Inst. Stat. Math. 2:61–67.
BAKER, G. A. and G. C. HANNA
1949. Transformations of split-plot yield trial data to improve analysis of variance. Proc. Amer. Soc. Hort. Sci. 53:273–75.
BAKER, G. A., M. R. HUBERTY and F. J. VEIHMEYER
1952. A uniformity trial on unirrigated barley of ten years duration. Agron. J. 44:267–70.
BAKER, G. A. and E. B. ROESSLER
1957. Implications of a uniformity trial with small plots of wheat. Hilgardia 27:183–88.
BAKER, R. E. and G. A. BAKER
1950. Experimental design for studying resistance of strawberry varieties to verticillium wilt. Phytopathology 40:477–82.
HANNA, G. C. and G. A. BAKER
1951. Analysis of asparagus field trials on the basis of partial records. Proc. Amer. Soc. Hort. Sci. 57:273–76.
HOYLE, B. J.
1957. Stress indicators. Ann. Research Review of Agronomy Staff, Davis, January 28.
1958. Islands. Ann. Research Review of Agronomy Staff, Davis, January 28.
HOYLE, B. J. and G. A. BAKER
1959. A more efficient barley yield testing method. Barley News Letter 2:32–33.
1959. The SD technique and its use for conducting field trials. Mimeo., 4 pp.
1959. The analysis of field trials based on the concept of islands of variation. Agron. Abstracts, Amer. Soc. Agron. Cleveland Meetings, November 16–20.
1960. Factor analysis of twenty-eight independent field trials on nine strains of Hannchen barley. Biometrics 16:127–128. Abstract 633.
1961. Game theory applied to field trials. Biometrics 17:167–68. Abstract 693.
RIDDLE, O. C. and G. A. BAKER
1944. Biases encountered in large-scale yield tests. Hilgardia 16:1–14.
SMITH, F. L.
1958. Effects of plot size, plot shape and number of replications on the efficiency of bean yield trials. Hilgardia 28:43–63.
WAYNICK, D. D.
1919. Variability in soils and its significance to past and future soil investigations. I. A study of nitrification in soils. Univ. of Calif. Pub. in Agric. Sci. 3:243–70.

WAYNICK, D. D. and L. T. SHARP
1949. Variability in soils and its significance to past and future soil investigations. II. Variations in nitrogen and in field soils and their relation to the accuracy of field trials. Univ. of Calif. Pub. in Agric. Sci. **4**:121–39.

WIEBE, G. A.
1935. Variation and correlation in grain yield among 1,500 wheat nursery plots. Jour. Agric. Research **50**:331–57.

4*m*-11,'64 (E8283) J.F. 141

The journal HILGARDIA is published at irregular intervals, in volumes of about 650 to 700 pages. The number of issues per volume varies.

Single copies of any issue may be obtained free, as long as the supply lasts; please request by volume and issue number from:

Agricultural Publications
University Hall
University of California
Berkeley, California 94720

The limit to nonresidents of California is 10 separate titles. The limit to California residents is 20 separate titles.

The journal will be sent regularly to libraries, schools, or institutions in one of the following ways:

1. In exchange for similar published material on research.
2. As a gift to qualified repository libraries only.
3. On a subscription basis—$7.50 a year paid in advance. All subscriptions will be started with the first number issued during a calendar year. Subscribers starting during any given year will be sent back numbers to the first of that year and will be billed for the ensuing year the following January. Make checks or money orders payable to The Regents of The University of California; send payment with order to Agricultural Publications at above address.