Biometrika (1975), **62**, 3, *p.* 665
With 3 text-figures
Printed in Great Britain

A comparison of different criteria for selecting binary tests in diagnostic keys

BY J. C. GOWER AND R. W. PAYNE
Rothamsted Experimental Station, Harpenden, Hertfordshire

SUMMARY

The problem of selecting tests to be used in nonprobabilistic binary diagnostic keys is discussed. Five selection criteria are compared and it is shown that all except a new criterion suffer from some deficiency. This criterion cannot be extended easily to cope with multi-response tests but another criterion, which behaves satisfactorily with binary tests, can be extended in some circumstances. The only criterion which can always be used with multi-response tests is least satisfactory for binary tests.

Some key words: Diagnostic keys; Error rates; Test selection.

1. INTRODUCTION

Consider n populations characterized by p binary characters. An individual sample from one of the populations may be assigned to its population of origin, or identified, by observing the values of its characters. In the following this process of observation will be termed testing and, unless otherwise stated, each test will be assumed to have two possible responses, termed positive and negative. A binary diagnostic key is a device for identifying samples, by applying the tests sequentially in a hierarchical manner. A sequence of test responses leading to an identification is termed a branch of the key. In general, branches will differ in length and the tests they use, although the same test can occur on several branches. Voss (1952) gives an interesting review of the historical development of keys.

To construct a key one requires a table giving the responses to the p tests for every population. There are four possible entries in this table: the response may be known to be positive for the whole population, or to be negative, or to be variable within the population, or the response may be unknown, there being no information available about the population value of the character concerned. Clearly, in the latter case, samples from the population may give either positive responses or negative responses and different samples need not give the same response. The first two types of response enable populations to be separated with certainty and when sufficient of these so-called fixed responses occur, all n populations can be identified uniquely. Even when there are insufficient fixed responses to give complete separation of the populations, one may identify a sample to within a group of populations with certainty. Such groups may be further separated, either by using a probabilistic key (Good, 1970) or better by discriminant methods. In this paper we are concerned solely with certain identification and therefore regard variable and unknown responses as equally uninformative, referring to both as unknown responses.

An optimum key may be defined as that with minimum average number of tests for identification. Alternative formulations include keys with minimum cost per identification or minimum number of different tests used (Gower & Barnett, 1971; Willcox & Lapage,

1972). Except for dynamic programming algorithms, which effectively enumerate all possible keys (Garey, 1972), no exact algorithm is known for finding optimum keys. These are impracticable for most real data which may be concerned with several hundred populations and up to about one hundred characters (Barnett & Pankhurst, 1974). Several authors (Pankhurst, 1970; Hall, 1970; Morse, 1971; Payne, 1974) present algorithms giving approximate solutions. These all operate by selecting first the test that best divides all the populations into two sets. Various criteria, some of which are described below, have been used to define what is meant by the best test. After the first division, the chosen criterion is used to select the next test to be used with each subset of populations, and so on. Garey & Graham (1974) give examples showing that selecting tests in this way, without examining their later consequences, can lead to inefficient keys, but most authors claim that their algorithms work well in practice and certainly give keys as good as, if not better than, those prepared by intuitive methods.

This paper compares and contrasts three criteria used to determine the best test and also discusses two further criteria, one of which is new. When all responses are fixed, these criteria all select the test that most nearly divides the populations into two equal groups. When some populations have unknown responses to a test, the populations concerned must be allocated to both groups. The criteria then differ but aim to select tests with few unknown responses while keeping the group sizes comparable.

We shall define p_i to be the proportion of populations giving positive responses to the ith test, q_i to be the proportion of populations giving negative responses to the ith test, and r_i to be the proportion of populations giving unknown responses to the ith test.

Thus, $p_i+q_i+r_i = 1$ and the ith test may be represented as a point T_i in barycentric coordinates (Gower, 1967, p. 23). Thus in Fig. 1, the midpoint of PQ, labelled O, represents the best possible test and R the worst possible test. Generally the better tests are those in the region of O.

It may sometimes be preferable to identify some populations more readily than others, either because they occur frequently or because they are more important in some sense. In these circumstances one may ascribe a weight to each population and redefine p_i, q_i and r_i as weighted proportions. In particular these weights might be prior probabilities for the populations.

2. THE CRITERIA AND SOME PROPERTIES

We considered three criteria in detail: DV (Morse, 1971), S (Seshu, 1965; Gower & Barnett, 1971), and C (Gower & Barnett, 1971).

$$\text{DV}_i = -2p_iq_i - \tfrac{1}{2}r_i(p_i+q_i), \tag{1}$$

$$S_i = (p_i+r_i)\log(p_i+r_i) + (q_i+r_i)\log(q_i+r_i), \tag{2}$$

$$C_i = (p_i-\tfrac{1}{2})^2 + (q_i-\tfrac{1}{2})^2 + r_i^2. \tag{3}$$

To ensure that all three criteria are to be minimized, DV is here defined as the negative of the criterion proposed by Morse (1971). The left-hand side of Fig. 1 shows contours at equal intervals for each criterion. Tests lying on the same contour are regarded as equivalent. Gower & Barnett (1971) suggests that tests lying within a band around a contour might also be regarded as equivalent. When selecting one from several equivalent tests, additional considerations can be taken into account. For example, one can select the test with least

cost, or the test that has been most used on other partially, or wholly, constructed branches of the key, or the test with fewest unknown responses. The band is merely the region between contours with criterion values differing by a small constant, similar to the wider bands shown on the left-hand side of Fig. 1. The bandwidth is governed by the rate of change of the criterion.

The criterion DV has parabolic contours and C has circular contours. All three have roughly circular contours near O. On the line PQ, where $r = 0$; $\text{DV} = 2(p-\frac{1}{2})^2-\frac{1}{2}$ and

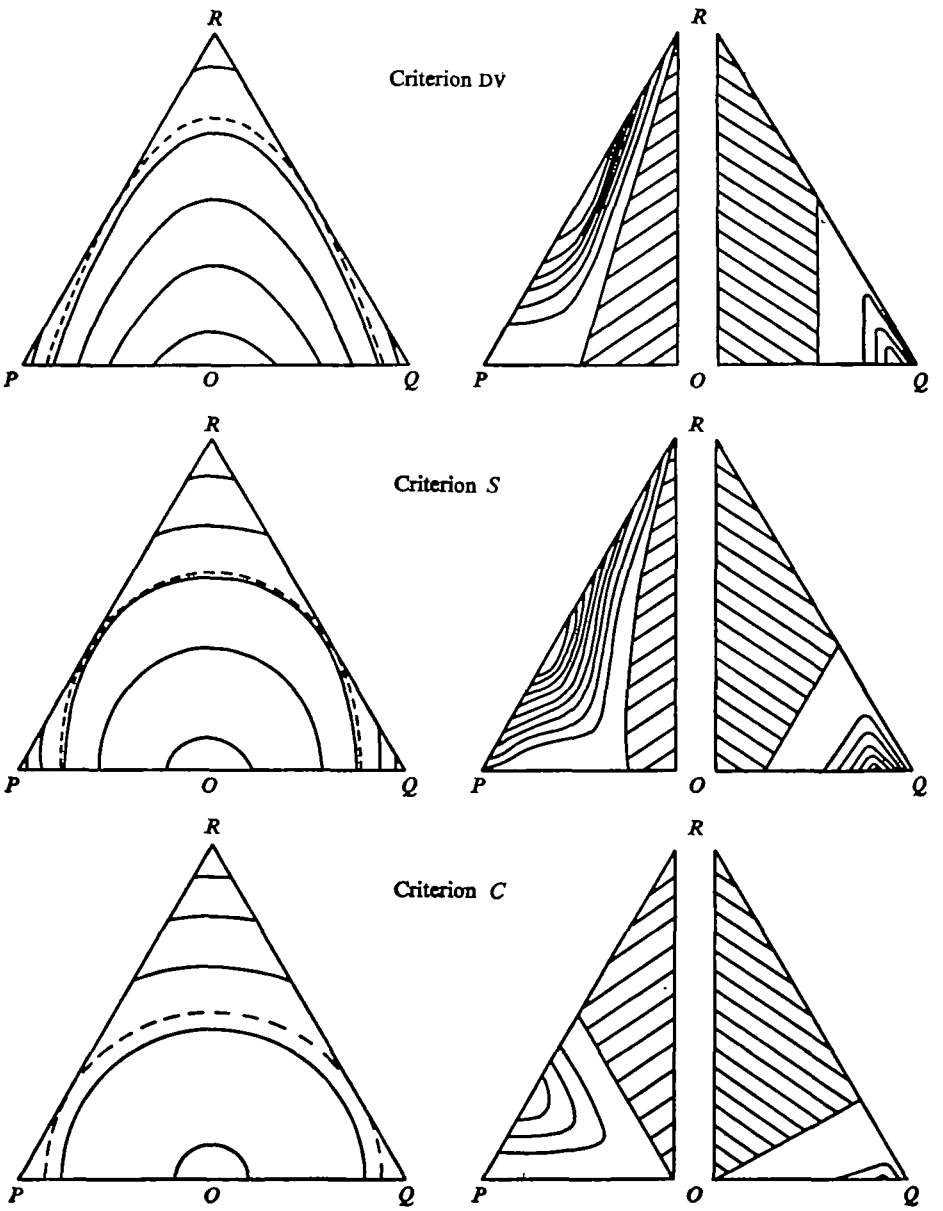

Fig. 1. Left-hand column: triangles show contours at equal intervals for each criterion. Right-hand column: triangles on the left, contours of type I error at intervals of 0·005; triangles, on the right, contours of type II error at intervals of 0·03. Shaded areas: regions with no error. Dotted line: contours tangential to PR and QR.

$C = 2(p-\frac{1}{2})^2$ differ only by a constant, and S behaves similarly to DV but its value initially increases less steeply from the minimum at O. On the lines $p-q=k$, that is lines perpendicular to PQ, DV increases linearly with r and $C = \frac{1}{2}(k^2+3r^2)$; S behaves similarly to C increasing less steeply from $p=\frac{1}{2}$.

Definite separation of populations cannot be guaranteed when either p or q is zero, and therefore such tests, termed indefinite below, should only be used when there are no alternatives. Yet the contours for all three criteria cut the lines PR and QR. The bigger the areas of the contours that touch the lines PR and QR the fewer indefinite tests will be selected in preference to tests giving definite separation. Tangential contours are shown as dotted lines in Fig. 1. When $p=0$ the contour for C touches QR at $q=0{\cdot}75$, $r=0{\cdot}25$, the contour for DV at $q=0{\cdot}5$, $r=0{\cdot}3$ and the contour for S at $q=1-1/e=0{\cdot}6321$, $r=1/e=0{\cdot}3679$. It is clear that DV encloses the greatest area and that S encloses slightly more than C. This implies that DV is less susceptible to accepting indefinite tests than are S and C. The criteria could be modified to reject tests lying on the boundaries, but tests near the boundaries should also be excluded because they are inferior to tests near P and Q with greater criterion value; see the discussion of type I error below.

3. ERRORS ASSOCIATED WITH THE CRITERIA

In some situations it is clear that one test is better or worse than another. If for two tests T_i and T_j

$$\max(p_i, q_i) > \max(p_j, q_j), \quad \min(p_i, q_i) > \min(p_j, q_j), \tag{4}$$

then T_i is certainly a better test than T_j, because more samples will be assigned definitely to both the positive and the negative response groups. Conversely if

$$\max(p_i, q_i) < \max(p_j, q_j), \quad \min(p_i, q_i) < \min(p_j, q_j), \tag{5}$$

then T_i is certainly a poorer test than T_j. Equations (4) and (5) are expressed in terms of $\max(p_i, q_i)$ and $\min(p_i, q_i)$ because for any test, an equivalent test can be defined which interchanges the definitions of positive and negative responses. In the following we assume that this relabelling has been done to make $p_i \geqslant q_i$ and $p_j \geqslant q_j$. Thus we only consider the region to the left of OR. Regions containing tests better and worse than a given test T are shown on the left-hand side of Fig. 2. Tests in the unshaded regions are not clearly better or worse than T.

Using any of the criteria to compare T with other tests, two types of error may arise. A type I error occurs when the criterion defines a test to be poorer than T, though (4) shows that it is better, hence it is possible for T to be selected in preference to other, better, tests. A type II error occurs when the criterion defines a test to be better than T though (5) shows that it is worse. Thus, if there were no better tests available, the criterion might select a test worse than T in preference to T. No criterion with convex contours can have tests with type II errors only, except on the boundary PQ or if they contain sections parallel to PR; see criterion BP, below. On the right-hand side of Fig. 2, tests T_1, T_2 and T_3 are, respectively, tests with both type I and type II errors, type I error only, and no errors. For the criteria DV, S and C the areas associated with both errors have been calculated for a range of tests. As these errors are symmetric about OR, contours of equal error have been shown for both types in the pairs of triangles on the right-hand side of Fig. 1. The left-hand triangle gives the type I errors and the right-hand triangle the type II errors. The contours are at the same

intervals for the three criteria; for type I errors at step lengths of 0·005 and for type II errors at step lengths of 0·03, taking the area of PQR as $\sqrt{3}$. The circle criterion C is least prone to both types of error followed by DV, while S is poorest in both respects. Although DV and S have zero type I and type II errors near O, the most useful region, C has only negligible error in this region.

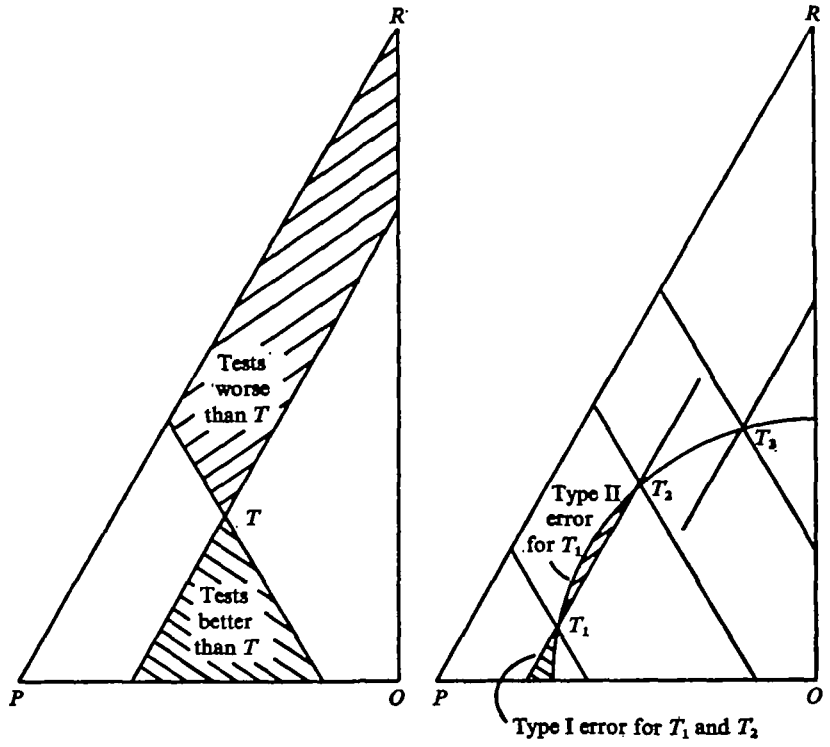

Fig. 2. Left-hand triangle: regions containing tests definitely better and definitely worse than T. Right-hand triangle: test T_1 has both type I and type II errors, test T_2 type I error only and test T_3 no error.

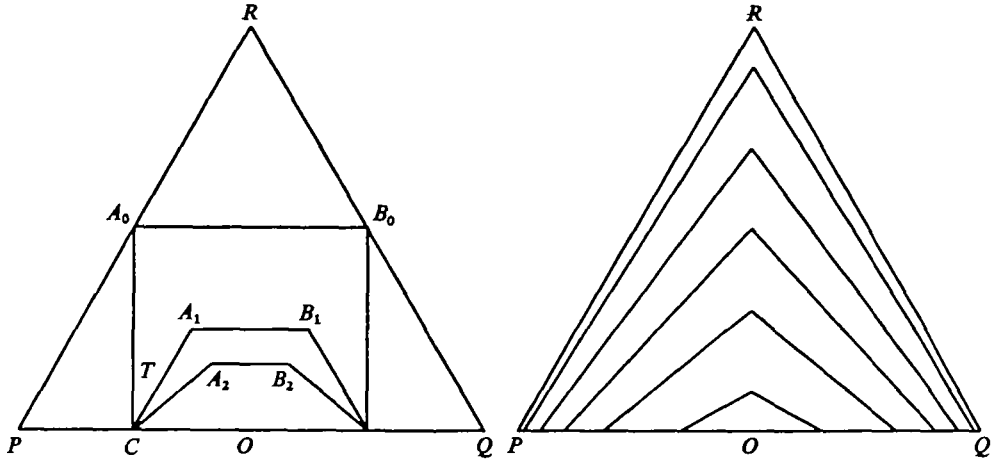

Fig. 3. Left-hand triangle: contours of BP through $p = \frac{1}{2}$, $q = \frac{1}{2}$ for $k = 0$, 1 and 2. Right-hand triangle: contours at equal intervals of GP with $\theta = \frac{1}{2}$.

Criteria with no type I or type II errors can be defined. Barnett & Pankhurst (1974) use

$$\text{BP}_i = |p_i - \tfrac{1}{2}| + |q_i - \tfrac{1}{2}| + kr_i,$$

where $k = n$, but we shall consider other nonnegative integer values of k. Figure 3 shows contours through the point $p = \frac{1}{4}, q = \frac{3}{4}$ for $k = 0$, 1 and 2. When $k < 1$ type I and II errors occur; when $k = 1$ a test T on A_1C has vestigial type II error on the line A_1T. When $k > 1$ neither type I nor type II errors occur. However, tests on any of the lines A_jB_j $(j = 0, 1, 2)$ are considered equivalent, whereas tests on OR are clearly the best of such groups.

Contours which are isosceles triangles based on PQ avoid this deficiency and will have zero type I and type II error, provided their base angles are less than $2\pi/3$. By increasing the base angle as the triangles increase in size, contours may be prevented from cutting the lines PR and QR, so avoiding indefinite tests. The following criterion fulfils all these conditions:

$$\text{GP}_i = -\min(p_i, q_i)/[\theta + (1-\theta)\{r_i + 2\min(p_i, q_i)\}], \tag{6}$$

where $0 \leqslant \theta < 1$. As the triangular contours increase in size their slopes increase from $\theta\sqrt{3}$ to $\sqrt{3}$ as shown on the right-hand side of Fig. 3. Choice of θ governs how heavily one wishes to guard against using unknown responses. High values of θ give contours which include more unknown responses than those for low values of θ. Values of θ near zero will not distinguish very well between tests on PQ, culminating in regarding all these as equivalent when $\theta = 0$. If $\theta = 1$ were allowed, (6) would become $\text{GP}_i = -\min(p_i, q_i)$, which gives nested equilateral triangle contours, corresponding to (4) and (5), which have type II errors.

4. TESTS WITH MORE THAN TWO RESPONSES

It is always possible to recode a test with m responses as $m - 1$ binary pseudotests. This is often sufficient but may be unsatisfactory when costs of tests are significant. For example, suppose the n populations can be separated by a single test with n responses. When this test is regarded as $(n-1)$ pseudotests the criteria discussed above are unlikely to select just from the pseudotests and will therefore generally produce a more costly key. Thus methods are needed to assess multiresponse tests directly.

To compare a test with m_1 responses to one with $m_2 > m_1$ responses we note that they both may be regarded as dividing the populations into m_2 classes, but that with the former $m_2 - m_1$ of the classes are empty. When all responses are fixed we can choose $M = \max(m_i)$ and amend S_i and C_i to become

$$S_i^* = \sum_{j=1}^{m_i} p_{ij} \log p_{ij}$$

$$C_i^* = \sum_{j=1}^{M} \left(p_{ij} - \frac{1}{M}\right)^2 = \sum_{j=1}^{m_i} p_{ij}^2 - \frac{1}{M},$$

where p_{ij} is the proportion of populations in the jth class of the ith test. Thus both S_i^* and C_i^* rank the tests independently of the choice of M. It might seem that BP extends similarly to give

$$\text{BP}_i^* = \sum_{j=1}^{M} \left| p_{ij} - \frac{1}{M} \right|.$$

However, even with $M = 3$ the two sets of tests responses $(\frac{1}{2}, \frac{1}{2}, 0)$ and $(\frac{1}{3}, \frac{2}{3}, 0)$ both give $\text{BP}^* = \frac{1}{3}$ which does not preserve the ranking for $M = 2$. Thus this method of extension is

Table 1. *Performance of the five criteria*

Characteristic	Criterion				
	DV	S	C	BP	GP
Computation	Efficient	Inefficient	Efficient	Efficient	Efficient
Type I errors / Type II errors	Intermediate	Largest errors	Errors smaller than DV or S	None*	None
Prevention of indefinite tests	Good	Satisfactory	Satisfactory	Poor**	Perfect
Performance for r constant	Perfect	Perfect	Perfect	Poor	Perfect
Weighting of unknown responses	Fixed	Fixed	Fixed	Controllable***	Controllable
Multiresponse extension	No	Yes	Only when all responses fixed	Unsatisfactory	No

* Provided $k > 1$.
** Worsens with increasing k.
*** The choice of $k = n$ (Barnett & Pankhurst, 1974) seems excessively restrictive.

unacceptable. Pankhurst (1970) has an alternative method of extension which biases against tests with more than two responses.

With unknown responses we define

$$S_i^* = \sum_{j=1}^{m_i} (p_{ij} + r_i) \log (p_{ij} + r_i)$$

and S_i^* remains independent of M. However we have been unable to find any similar extension for C_i^* or BP_i^*. It might seem that a suitable extension of C_i^* would be

$$C_i^* = \sum_{j=1}^{M} \left(p_{ij} - \frac{1}{M}\right)^2 + r_i^2.$$

With $M = 3$, this would regard the two-response test (p_1, p_2, r) as equivalent to the three-response test $(p_1, p_2, 0, r)$ although the former separates the populations into groups with sizes proportional to $(p_1 + r, p_2 + r, 0)$ compared with $(p_1 + r, p_2 + r, r)$ for the latter. Thus the three-response test is better because it offers the possibility of separation, should samples actually occur with the third level of response.

Neither DV nor GP seem to extend simply to include tests with more than two responses.

5. Conclusion

Criteria for selecting binary tests with unknown responses, have been shown to have the following possible deficiencies: (i) they may produce type II and/or type I errors; (ii) they may select indefinite tests in preference to definite tests; (iii) for constant r, they may regard tests for which $p \neq q$, to be as good as a test with $p = q$. Apart from such clear deficiencies, there is the more subjective question of deciding how much weight should be given to unknown responses. The possibility of extending a criterion to deal with multiresponse tests is important for some applications. Finally, as the calculation of selection criteria occurs in the innermost loops of computer programs for generating diagnostic keys, their computa-

tional efficiency is important. Table 1 sets out the performance of the criteria in these respects.

For tests with only two responses GP is best on all counts, but C and DV are satisfactory. In the multiresponse case, only S seems to be suitable unless all responses are fixed.

REFERENCES

BARNETT, J. A. & PANKHURST, R. J. (1974). *A New Key to the Yeasts.* Amsterdam: North Holland.

GAREY, M. R. (1972). Optimal binary identification procedures. *S.I.A.M. J. Appl. Math.* **23**, 173–86.

GAREY, M. R. and GRAHAM, R. L. (1974). Performance bounds on the splitting algorithm for binary testing. *Acta Informatica* **3**, 347–55.

GOOD, I. J. (1970). Some statistical methods in machine intelligence research. *Math. Biosci.* **6**, 185–208.

GOWER, J. C. (1967). Multivariate analysis and multidimensional geometry. *Statistician* **17**, 13–28.

GOWER, J. C. & BARNETT, J. A. (1971). Selecting tests in diagnostic keys with unknown responses. *Nature* **232**, 491–3.

HALL, A. V. (1970). A computer-based system for forming identification keys. *Taxon.* **19**, 12–8.

MORSE, L. E. (1971). Specimen identification and key construction with time sharing computers. *Taxon* **20**, 269–82.

PANKHURST, R. J. (1970). A computer program for generating diagnostic keys. *Computer J.* **13**, 145–51.

PAYNE, R. W. (1974). Genkey: a program for constructing diagnostic keys. In *Biological Identification with Computers*, Ed. R. J. Pankhurst, pp. 65–72. London: Academic Press.

SESHU, S. (1965). On an improved diagnosis program. *I.E.E. Trans. Electronic Computers.* EC-14, 76–9.

VOSS, E. G. (1952). The history of keys and phylogenetic trees in systematic biology. *J. Sci. Labs. Denison Univ.* **43**, 1–25.

WILCOX, W. R. & LAPAGE, S. P. (1972). Automatic construction of diagnostic tables. *Computer J.* **15**, 263–7.

[*Received August* 1974. *Revised March* 1975]