DOCUMENT RESUME

ED 104 906 TM 004 355

AUTHOR	Larsson, Bernt
TITLE	The Stability of Results: Some Examples of the Effects of Scale Transformations. Didakometry, No. 42, 1974.
INSTITUTION	School of Education, Malmo (Sweden). Dept. of Educational and Psychological Research.
PUB DATE	Oct 74
NOTE	36p.; Paper read at the conference with Society of Multivariate Experimental Psychology (Frankfurt, Sept. 1974)
EDRS PRICE	MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS	*Analysis of Variance; *Correlation; Discriminant Analysis; *Factor Analysis; Mathematical Models; Matrices; Measurement Techniques; *Reliability; Sampling; *Statistical Analysis; Test Reliability
IDENTIFIERS	Binary Coding; Q Index; Scale Transformation

ABSTRACT

This report gives some simple examples of stability for one factor and 2 x 2 factorial analysis of variance, reliability and correlations. The findings are very different: from superstability (no transformation whatsoever can change the result) to almost total instability. This is followed by a discussion of applications to multivariate analysis, and by some final remarks. It can be added that the technique can also be utilized for scaling variables to obtain a best fit to mathematical models other than those involved in usual statistical analysis. (Author)

THE STABILITY OF RESULTS: SOME EXAMPLES OF THE EFFECTS OF SCALE TRANSFORMATIONS

Bernt Larsson

Larsson, B. The stability of results: Some examples of the effects of scale transformations. Didakometry (Malmö, Sweden: School of Education), No. 42, 1974.

When the admissible class of transformations for a scale; defined to measure a certain concept, is broader than the class of transformations for which a given index of result is invariant, the question of the stability of results arises. In such situations one may be interested in finding the range of the index or perhaps that transformation which maximizes or minimizes the index. The technique used here to obtain these objects is to express a variable with many categories as a weighted sum of its binary variables; the weights being the scale values.

This report gives some simple examples of stability for one factor and 2 x 2 factorial analysis of variance, reliability and correlations. The findings are very different: from superstability (no transformation whatsoever can change the result) to almost total instability. This is followed by a discussion of applications to multivariate analysis, and by some final remarks. It can be added that the technique can also be utilized for scaling variables to obtain a best fit to mathematical models other than those involved in usual statistical analysis.

Keywords: Measurement, transformations, scales

Paper read at the conference with Society of Multivariate Experimental Psychology in Frankfurt/M. in September 1974.

INTRODUCTION

Scales used in educational research are, as a rule, loosely defined. The most common numerical coding of the possible outcomes of a measurement is successive integers. However, there is seldom anything in the educational measurement procedure which prescribes this rather than any other coding. Educational researchers will in most cases not have any fundamental objection to exchanging this coding for a monotonic transformation of it.

On the other hand, many statistical methods (or other mathematical models used in educational research) are only invariant e. g. up to linear transformations. The question is: how stable are results, described by these methods, when monotonic transformations constitute the class of acceptable codings? High stability admits conclusions with great generality. It may also be of interest to choose that scale which, under given restrictions, maximizes (or minimizes) a certain index of result.

The techniques used for investigating the stability are based on a general principle. By using binary coding, each many-valued variable can be expressed as a weighted sum of its binary variables, where the weights are the scale values. This implies that almost all analysis will be multivariate, e.g. a certain type of analysis of variance (ANOVA) is transferred to the corresponding discriminant analysis, modified due to some restriction of transformations.

This report gives some simple examples of the stability of results for some statistical methods, viz. one factor and 2 x 2 facterial ANOVA with different cell samples, reliability estimates from a one factor ANOVA with repeated measures design, and product-moment correlations. The report also discusses, though without examples, some possible extensions to multivariate methods. With one exception, the examples only treat three or four-valued variables, which make it possible to visualize the results on graphs. The data are, again with one exception, artificial, constructed to constitute a first test of some optimization routines.

It can be added that the binary coding technique is not limited to statistical methods. We can use it to code variables, under given restrictions, to obtain a best fit to a certain mathematical model (described by a goodness-of-fit criterion chosen). If this optimal fit is bad, the conclusion that the model is unsuitable will be quite general. For instance, we may code a variable to obtain a certain distribution function, code two variables to obtain a given linear relation, or code a learning variable to be a specified function of the number of trials.

METHOD

In this section we will first describe the binary coding technique and relations between and within many-valued variables and binary variables. Some comments are then made on the general forms of the indices of result used in the examples, followed by a short discussion of the concept of stability and some simulations.

Binary coding

The idea of binary coding is not new. Some information about it is given in Larsson (1973) and Bradley et al. (1962) use it in a modified form. The description will here be sufficiently general to cover also most of the discussion concerning multivariate ~~analysis~~ analysis.

Let x_i, $i = 1, \ldots, p$, be a many-valued variable with $k_i + 1$ categories. There are n_{ig} measurement objects characterized by category g, which has the numerical code a_{ig}; $g = 0, 1, \ldots, k_i$. The categories are often ordered and in such cases g indicates the order. In the sequel we only regard a certain standardized coding having $a_{i0} = 0$ and $a_{ik_i} = 1$.

The binary variable u_{ig} is now defined as

$$\textcircled{1} \quad u_{ig} = \begin{cases} 1 & \text{if } x_i = a_{ig} \\ 0 & \text{if } x_i \neq a_{ig} \end{cases} \qquad g = 1, \ldots, k_i \, .$$

The vector of arithmetic means of the binary variables is $\mathbf{m}_i = \{n_{ig}/n\}$ and the covariance matrix is $\mathbf{S}_{ii} = \{\delta_{gh} n_{ig}/n - n_{ig} n_{ih}/n^2\}$, where δ_{gh} is Kronecker's δ and $n = \sum_{g=0}^{k_i} n_{ig}$. (We assume that n has the same value, independent of i .) Likewise, the covariance matrix between binary variables, corresponding to two x-variables, becomes $\mathbf{S}_{ij} = \{n_{g(i)h(j)}/n - n_{ig} n_{jh}/n^2\}$. Here h and j are alternative indices of g and i, respectively, and $n_{g(i)h(j)}$ is the number of objects which simultaneously belong to category g of x_i and category h of x_j.

As a parenthesis, we may mention that the nonnumerical information of x_i, e.g. that contained in $\mathbf{S}_{ii}$, can be used by analogue to multivariate statistics. The determinant of a covariance matrix is there one index of 'generalized variance' and $|\mathbf{S}_{ii}| = \prod_{g=0}^{k_i} n_{ig}/n^{k_i+1}$ may be used as a measure of the nonnumerical 'variance' of x_i. It is related to information theoretical measures of uncertainty, see e.g. Fhanér (1966).

For cases dealt with here, binary coding may be said to split up the information of x_i in a nonnumerical part, $\mathbf{u}_i = \{u_{ig}\}$, and a numerical part, $\mathbf{a}_i = \{a_{ig}\}$. We obtain the fundamental formula

2 $$x_i = \mathbf{a}_i'\mathbf{u}_i \;.$$

The arithmetic mean of x_i becomes $\mathbf{a}_i\mathbf{m}_i$ and its variance $\mathbf{a}_i'\mathbf{S}_{ii}\mathbf{a}_i$, while the covariance of x_i and x_j can be written as $\mathbf{a}_i'\mathbf{S}_{ij}\mathbf{a}_j$.

Let us now consider all x-variables simultaneously and define $\mathbf{x} = \{x_i\}$, $\mathbf{m}_x = \{\mathbf{a}_i'\mathbf{m}_i\}$, $\mathbf{u} = \{\mathbf{u}_i\}$, $\mathbf{m}_u = \{\mathbf{m}_i\}$ and $\mathbf{S}_{uu} = \{\mathbf{S}_{ij}\}$. We also need $\mathbf{D}$, a block diagonal matrix having $\mathbf{a}_i$ on the principal diagonal and thus of order K x p, where $K = \sum_1^p k_i$. Hence

3 $$\mathbf{x} = \mathbf{D}'\mathbf{u} \;.$$

It follows from formula 3 that $\mathbf{m}_x = \mathbf{D}\mathbf{m}_u$ and the covariance matrix of $\mathbf{x}$ will be $\mathbf{S}_{xx} = \mathbf{D}'\mathbf{S}_{uu}\mathbf{D}$.

In multivariate statistical analysis it is rather common to define new variables as a weighted sum of other variables, e.g. $z = \mathbf{c}'\mathbf{x}$. We may take $\mathbf{t} = \mathbf{D}\mathbf{c}$, meaning that $z = \mathbf{t}'\mathbf{u}$ with $m_z = \mathbf{t}'\mathbf{m}_u$ and $s_z^2 = \mathbf{t}'\mathbf{S}_{uu}\mathbf{t}$. Thus, the situation is the same as for one x-variable. (But see next section about formulations of restriction for monotonic transformations.)

Indices of result

Almost all indices of result presented in this report has the following form for one dependent variable x (we now skip i and j):

4 $$Q = \frac{\mathbf{a}'\mathbf{F}\mathbf{a}}{\mathbf{a}'\mathbf{G}\mathbf{a}}$$

Both matrices are real and symmetric, and they can be weighted sums of other, more basic matrices. For all Q here, **G** will be positive definite. There are, however, cases where **G** may be positive semidefinite, e.g. if Q is a F ratio for a random factor. In many cases **F** will be positive semidefinite (or definite) but we will also meet exceptions from this (**F** is indefinite). I think that exceptions are rather common in connection with variance components, where negative values are possible. In some applications there may be other properties, e.g. each diagonal element of **F** cannot exceed the corresponding element of **G**.

For standardized **a**, but no other restrictions, we can seek for an optimal scale in the whole (k-1)-dimensional real space of **a** and thus have an eigenvalue problem as e.g. for common discriminant analysis. The only restriction taken up here is that of monotonic transformations. In most cases this will mean $0 \leq a_1 \leq a_2 \leq \ldots \leq a_{k-1} \leq 1$. The admissible **a** space is then a peculiarly cut 'piece of cheese' in the principle quadrant. Under certain circumstances, however, monotonic transformations can only involve blockwise ranking, for instance $0 \leq (a_1, a_2) \leq a_3 \leq 1$, with no ranking within blocks. Such a case will appear in this report. Also, for many x-variables the monotonic restriction implies that the **t** vector of the last section will only be ranked within blocks ($c_i \mathbf{a}_i$) but not between blocks.

I believe that it is not unusual for optimal **a** to lie on the boundary of the admissible space. In particular, corner solutions seem to be 'favoured' for min Q, as far as my brief experience hitherto shows, that is, x is dichotomized. Some support for this belief concerning max Q is given by Bradley et al. (1962). They seem to have analysed rather a lot of data (one factor ANOVA) and often found boundary solutions, at least when k is large.

The index Q according to formula 4 is not relevant for one of the examples concerning a productmoment correlation. The problem is then simultaneously to code two x-variables and Q will have the general form.

$$5 \qquad Q = \frac{(\mathbf{a}_i' \mathbf{F}_{ij} \mathbf{a}_j)^2}{\mathbf{a}_i' \mathbf{G}_{ii} \mathbf{a}_i \mathbf{a}_j' \mathbf{H}_{jj} \mathbf{a}_j}$$

For the (squared) correlation, the matrices are different covariance matrices (between and within $\mathbf{u}_i$ and $\mathbf{u}_j$).

Stability

For a certain Q-index and given restrictions, data are more stable, or more insensitive to admissible transformations, the lesser the difference between the maximum and minimum of Q. We will not use any special stability measure in this report but it can be needed for certain comparisons. There are Q-indices, the range of which vary (e.g. as a function of n), and different Q-indices may have quite different ranges. For indices with finite ranges it is reasonable to relate the actual range (max Q - min Q for certain data) to the maximally possible range (without restrictions), e.g. define stability as 1 - (max Q - min Q for certain data)/(maximal range).

Total instability is obtained for data which have maximal Q-range. Som examples have data which are almost in this state. The opposite will be coined superstability, which means that no transformation - monotonic or not - can change Q. For formula 4 this implies that **F** is proportional to **G**. We will give two examples of this remarkable property. It is finally obvious, for a definition of stability as of the last paragraph, that for two different restrictions, described by $\mathbf{a} \in R_1$ and $\mathbf{a} \in R_2$ with $R_1 \subset R_2$, the stability cannot be greater for R_2 than for R_1.

Simulations

Two types of simulations will be commented upon here, but only the first type has yet been performed. The type I subroutine produces rectangularly distributed random scale values which are ranked and exploited for the calculation of Q according to formula 4 (**F** and **G** are fixed and supplied by the main program). The generation of **a** is repeated an arbitrarily number of times, thus giving a whole distribution of Q. It is of special interest to know the relative position of Q for equally spaced scale values. (You may here speak about a kind of inference with the generated distribution as a sample distribution over scales.) The type I runs will also serve as a check of the optimization routines: if the simulated distribution contains more extreme values than those from the optimization routines, an error is indicated.

The purpose with type II simulations is to get a comprehension of the variation of the extreme values with repeated samples of measurement objects from the same population. We will construct some convenient populations, take, a number of samples and apply the optimization subroutines. In this way we get an estimate of the common kind of sample distribution of min Q and max Q, which is, no doubt, important. However, this type of analysis seems to be rather expensive and cannot always be made. I assume that some priority must be made: it may be necessary to elucidate this problem by only running type II simulations for the most common Q-indices.

SOME EXAMPLES

Most of the following examples are illustrated both with tables of basic data and with graphs on Q as a function of **a**. While the tables are presented on successive text pages, the graphs are collected in an appendix. The matrices **F**, **G** and **H** are not shown but they are easily retrieved from the appendix, where the functions are also given.

Two simple ANOVA designs

Factorial ANOVA with different cell samples has been studied by some authors (not all referred to here but see Meredith, Fredriksen & McLaughlin, 1974, for some further references) aiming at finding scales which optimize a certain effect. Tukey (1950) is one of the first to solve, at least partially, this problem. He maximizes the F ratio but his method does not guarantee rank invariance. Box & Cox (1964) give this problem a more complete solution, but they restrict themselves to certain families of functions. In that respect the method described by Kruskal (1965) and Kruskal & Carmone (1969) is more general: it considers all functions within the class of monotonic transformations. This is also the case with the method proposed by Bradley et al. (1962) and in this report.

One factor ANOVA

For a univariate one factor ANOVA with different samples, the total sum of squares is divided up into the sum of squares between groups (samples) and within groups. The corresponding cross product matrices in the multivariate case will be denoted **T**, **B** and **W**, respectively. We generate these matrices, of order k x k, by binary coding of a dependent variable with k + 1 categories. The Q index used here, for given scale values **a** , will be the ratio of the sum of squares between groups to the total sum of squares. In accordance with formula 4 this implies that **F** = **B** and **G** = **T**.

The first example is taken from Larsson (1973). As is clear from table 1, the factor has three levels and the dependent variable three categories. (The numbering of the latter only indicates order.) Figure 1 of the appendix shows Q as a function of **a** .

Table 1. Basic data of example 1

	A_1	A_2	A_3	Σ
3	0	10	10	20
2	1	29	20	50
1	29	1	0	30
Σ	30	40	30	100

The two eigenvalues become 0.9079 and 0.0069, of which the largest one happens to be generated by an admissible scale under the restriction of monotonic transformations. The minimum of Q with this restriction is 0.1146. It can be added that the scale (0.0, 0.5, 1.0) gives a Q value of 0.6610. We thus have a very instable situation, where different monotonic transformations may generate quite different descriptions: the proportions of the total variance explained by group differences may differ as much as 79 %. Notice also that the dichotomized scale (0.0, 1.0, 1.0) is more sensitive to group discrimination than (0.0, 0.5, 1.0). I believe that this can be a rather-general finding: more scale values do not guarantee higher Q values.

The basic data of the next example are shown in table 2. It consists of two parts, each with two levels and a three-valued dependent variable. Figure 2 of the appendix gives both curves (Q as a function of **a**).

Table 2. Basic data of example 2

	A_1	A_2	Σ	A_1	A_2	Σ
3	10	25	35	0	30	30
2	20	10	30	40	0	40
1	10	25	35	0	30	30
Σ	40	60	100	40	60	100

For both parts, data are constructed so that the scale (0.0, 0.5, 1.0) gives a Q value of 0.0000. The left part has eigenvalues of 0.1270 and 0.0000 and the maximal Q value for monotonic transformations is 0.0293. The right part, which involves extremely different distributions, has eigenvalues of 1.0000 and 0.0000, while the restricted Q maximum is 0.2857. Thus, a Q value of zero for equally spaced scale values can be increased, though very dissimilar distributions seem to be needed for a substantial change. If the distributions have exactly the same form, Q will be superstable (will be zero independent of **a**).

2 x 2 factorial ANOVA

For this case the crossproduct matrix between cells will be partioned into three matrices: $\mathbf{B}_A$ for the main effect of factor A, $\mathbf{B}_B$ for the main effect of factor B, and $\mathbf{B}_{AB}$ for the interaction effect. We use the same Q index as for one factor ANOVA, which means that the numerator matrix of formula 4 if one of the **B** matrices, while **G** is still equal to **T**. When we describe an effect by this Q value, it is evident that the effect can be totally eliminated in the numerator matrix is positive semidefinite. This property is normally obtained when the degree of freedom of the effect is less than k. However, it is far from certain that a monotonic transformation gives Q = 0.

Two different examples will be given for the 2 x 2 factorial design with independent cell samples. The basic data of the first one is presented in table 3. Figure 3 of the appendix shows the curves of the effects, including that between cells.

Table 3. Basic data of example 3

	A_1		A_2		
	B_1	B_2	B_1	B_2	Σ
3	5	0	10	15	30
2	15	10	5	10	40
1	5	15	10	0	30
Σ	25	25	25	25	100

For equally spaced scale values we get $Q_A = Q_{AB} = 0.1500$ and $Q_B = 0.0000$. The B effect is an instance of superstability, Q_B is constantly zero, and its curve in figure 3 is not apparant as it coincides with the horisontal axis. The eigenvalues of A and AB are both 0.1917 and 0.0000, of which the highest one is associated with the admissible **a** space for monotonic transformations. With this restriction the minimal

Q value is 0.0476 for both effects. (Notice from figure 3 that the A and AB curves are reflections of each other around a = 0.5.) The sum effect (between cells) has eigenvalues 0.3000 and 0.0833. Here again the global maximum comes from the admissible **a** space but its minimum is 0.2381. The Q value between cells is quite stable for monotonic transformations, while that for A and AB is not quite so stable.

We have said that whenever the distributions of different groups are identical, Q = 0 is a superstable result. However, superstability is not confined to zero effects, as will be shown by the next example. The basic data for this can be found in table 4, and figure 4 of the appendix illustrates the functions.

Table 4. Basic data of example 4

	A_1		A_2		
	B_1	B_2	B_1	B_2	Σ
3	10	40	10	20	80
2	40	10	10	20	80
1	10	10	40	20	80
Σ	60	60	60	60	240

For the usual scale (0.0, 0.5, 1.0) we obtain $Q_A = Q_B = 0.0938$ and $Q_{AB} = 0.0000$. The eigenvalues are 0.1250 and 0.0000 for all three effects. When restricting ourselves to monotonic transformations, the restricted maxima and minima are 0.1250 and 0.0313 for A and B, while those for AB are 0.0313 and 0.0000. However, the remarkable property of this example is $Q_A + Q_B + Q_{AB} = 0.1875$, irrespective of the scale. No transformation whatsoever can change the proportion of the total variance due to differences of the cell means. I have no idea whether data which, at least roughly, have this property are common or not. Notice that the concept of superstability can be dependent on Q: it is not certain that an index describes a result as superstable, in spite of the fact that it has been so described by another index.

There are several conceivable indices suitable for describing ANOVA results. Besides the proportion already used, we may mention the F-ratic and different combinations of variance components. As an example of an alternative index, we take $Q = \hat{\sigma}^2_A / \hat{\sigma}^2_e$, the ratio of the estimated variance component of factor A to the corresponding component of error, and apply this index to the first 2 x 2 factorial example.

We assume that A and B are both fixed and estimate the components by equating the observed mean squares with their expected values. The index has the form shown by formula 4 with $\mathbf{F} = (\mathbf{B}_A - \mathbf{W}/96) / 50$ and $\mathbf{G} = \mathbf{W}/96$. Its lower limit is thus dependent of data (here $-1/50$), while the upper limit may be set to infinity. The eigenvalues are, for this example, 0.4786 och -0.0200. For monotonic transformations, $0.1000 \leq Q \leq 0.4786$. As is seen from figure 5, the curve for this index bears a close resemblance to the Q_A curve of figure 3. This may, however, be a mere coincidence. For instance, if we take the same index but assume the factors to be random, the resulting curve is rather different from the Q_A curve of figure 3.

Reliability

Determination of a weighted sum of variables with maximal reliability is by no means a new problem. One of the older methods is presented e.g. in Lord & Novick (1968, pp. 123-124) and another more general method is described by Abelson (1960). These methods work with the same form of Q (see formula 4) as the method proposed here, but the matrices are not the same. Besides, my method can guarantee a solution within the class of monotonic transformations and can be used for a single variable. This is not the case with the other two methods.

We shall take an example which admits a comparison with Abelson's method. The example comprises a 'test', composed by two binary items, which is measured on ten persons on two occasions. The basic data are given in table 5 and a Q function in figure 6.

Table 5. Basic data of example 5

		Item 1		Item 2	
		Occasion		Occasion	
		1	2	1	2
	1	1	1	1	1
	2	1	0	1	1
	3	1	1	1	1
	4	1	1	0	0
Person	5	1	1	0	0
	6	0	1	1	0
	7	0	0	1	0
	8	0	0	1	0
	9	0	0	0	1
	10	0	0	0	0

A '1' may be interpreted as a correct answer, a '0' as a wrong answer. The test has four possible outcomes: (0, 0), (0, 1), (1, 0) and (1, 1) with scale values 0, a_1, a_2 and 1, respectively. The restriction to monotonic transformations will here imply $0 \leq a_1, a_2 \leq 1$, since there is no clear way of internally ranking the outcomes (0, 1) and (1, 0). Table 5 corresponds to a one factor ANOVA with repeated measures for a given scale. In such a design the total sum of squares is split up into three sums: for occasions O, persons P and interaction (plus error) OP, and the same split is valid for the cross product matrices: $\mathbf{T} = \mathbf{B}_O + \mathbf{B}_P + \mathbf{B}_{OP}$. The estimates of variance components relevant to reliability give, in this case, $\mathbf{F} = \mathbf{B}_P - \mathbf{B}_{OP}$ and $\mathbf{G} = \mathbf{B}_P + \mathbf{B}_{OP}$.

The eigenvalues are 0.8383, 0.6762 and -0.4645, none of which corresponds to the admissible **a** space for monotonic transformations. The common scale (0.0, 0.5, 0.5; 1.0) gives a reliability of 0.6327. If we do not differentiate between the outcomes (0, 1) and (1, 0), this value can still be improved on: the scale (0.0, 0.15, 0.15, 1.0) has a reliability of 0.7747. (It happens to be the larger eigenvalue for the Q function with $a_1 = a_2$.) According to Spearman-Brown's formula, this is equal to a doubled test with the common scale.

Abelson's method distinguishes between (0, 1) and (1, 0) but the solution is, for this example, confined to the line $a_1 + a_2 = 1$. It gives the reliability value 0.6939, which corresponds to 1.32 times the length of the commonly scaled test. (It seems to me that Abelson's method coincides with mine when the outcome of the items is reproducible from the sum score.) However, best of all monotonic transformations is (0.0, 0.2, 0.0, 1.0), generating the value 0.8029 (2.37 times the length of the common test). Finally, it can be said that the situation is unstable: the minimum value is 0.0875 when (0, 1) and (1, 0) are separated and 0.2162 without this separation.

The example presented above can be generalized to more complex univariate designs, e.g. those described by Cronbach et al. (1972). As far as I can see this involves no mathematical novelties: it is only a matter of correctly choosing the weighted sums of basic matrices (from the ANOVA) which define $\mathbf{F}$ and $\mathbf{G}$.

Correlations

The usual stability analysis of a squared bivariate correlation involves a Q index according to formula 5. As we shall see, however, there are, occasions when a correlation takes the form hitherto discussed for Q.

The first example makes use of data from table 1, where we now also assume the levels to be ordered (according to index numbers). Figure 7 of the appendix shows the Q function. The eigenvalues are the same as for example 1, 0.9079 and 0.0069, with the largest one coming from the admissible a space for monotonic transformations. There are nonmonotonic transformations for which Q becomes zero, but for monotonic transformations the minimal Q value is 0.0476. The scale (0.0, 0.5, 1.0) for both variables gives a value of 0.5173. This example thus shows a very unstable situation.

Figure 7 also gives some relations between this example and example 1. The curve denoted P_1 is connected to figure 1: the curve of figure 1 shows the height when following P_1, of figure 7. In the same manner we get P_2, the corresponding curve to P_1 when independent and dependent variables change places in example 1. (Only parts of P_1 and P_2 are shown in figure 7.)

Suppose that it is desirable to determine the same scale for a number of variables with equally many categories and to define an average correlation as the ratio of the average covariance to the average variance. (The reliability estimate of example 5 is such a correlation.) We then have a correlation analysis where the form of Q is given by formula 4. No example of such an analysis is shown here, but we will instead present data of another correlation problem conformable to formula 4.

This example comprises 'real' data from a pilot study (n = 44). The correlation problem concerns the relation between frequency statements (the number of days per year) and verbal statements for six different questions. The verbally anchored variables have categories labelled almost never, seldom, sometimes, often and almost always. The six questions asked refer to how often you 1. watch TV, 2. go to the pictures, 3. wake up rested, 4. have a headache, 5. are stressed and 6. feel expectant. Some correlations between frequency statements and verbal statements are given in table 6 for each question.

Table 6. Some correlations of example 7

	min Q	common Q	max Q
1	0.2921	0.7526	0.7553
2	0.1464	0.4570	0.5319
3	0.1212	0.7194	0.7447
4	0.0650	0.6065	0.6268
5	0.0950	0.4007	0.4212
6	0.0729	0.5838	0.6891

It is reasonable, for this example, not to recode the frequency variables: we regard 'the number of days per year' as a fixed scale and are only interested in numerically coding the verbal categories to obtain minimal and maximal squared correlations between frequency statements and verbal statements. This problem gives a Q index according to formula 4, with $\mathbf{F} = \mathbf{s}\,\mathbf{s}'/s^2$ and $\mathbf{G} = \mathbf{S}$. Here $\mathbf{s}$ is the vector of covariances between the binary variables and the frequency variable, which has variance s^2, and $\mathbf{S}$ is the covariance matrix of the binary variables.

As $\mathbf{F}$ is positive semidefinite, it is possible to obtain zero correlations but they do not correspond to scales within the $\mathbf{a}$ space of monotonic transformations. The minimal Q values for this space are all generated by corner solutions, that is, the worst admissible dichotomizations. The restricted maximal Q value coincides with the greatest eigenvalue, except in questions 2 and 3, which give the only boundary solutions, but their maximal Q values are almost the same as their nonzero eigenvalues. The second column of table 6 refers to Q when the verbal scale has equally spaced values. We see that these common Q values are of the same magnitudes as the corresponding maximal values, perhaps with the exception of question 6. On the other hand, the ability to predict frequency statements from verbal statements is in no case very high.

EXTENSIONS

This section contains some rather loose ideas about possible applications of stability analysis to multivariate statistical methods. I do not know if there are new numerical problems not encountered in univariate analysis.

Of course, the multitude of values to determine may in itself raise difficulties. I will now comment superficially upon principal component analysis, discriminant analysis, canonical correlation analysis and factor analysis.

We discuss principal component analysis only by treating the problem of finding a weighted variable $z = \mathbf{c}'\mathbf{x} = \mathbf{c}'\mathbf{D}'\mathbf{u} = \mathbf{t}'\mathbf{u}$ with maximal variance. For instance, for two variables with three categories each, we have $\mathbf{t}' = (c_1a_{11}, c_1, c_2a_{21}, c_2)$. The usual restrictions imply that $\mathbf{a}_i$, will be seperately ranked and that $\mathbf{c}'\mathbf{c} = 1$. However, I imagine that there will often be more restrictions. To use Jöreskog's words, see e.g. Jöreskog (1973), every element of t can be fixed, constrained or free.

Some variables, like the frequency variable of example 7, may be so well defined that its scale vector is fixed. Another instance of fixed values is to predetermine $\mathbf{c}$: you have a model about how z should be defined and investigate whether the best scaling reaches a sufficiently high variance. If you are not satisfied with the resulting variance then your model is not good under any monotonic transformations. In case some or all variables have the same number of categories it may be desirable to let the scale vectors be identical. This is a reasonable example of constrained values. Under some combination of fixed, constrained and free elements of $\mathbf{t}$ one is now interested in determining $\mathbf{t}$ such that $\max_{\mathbf{t}} \mathbf{t}'\mathbf{S}_{uu}\mathbf{t}$ (and perhaps also $\min_{\mathbf{D}} \max_{\mathbf{c}} \mathbf{t}'\mathbf{S}_{uu}\mathbf{t}$) is obtained.

Discriminant analysis is illustrated by finding the best discriminant function for a one factor design with independently sampled groups. Let $\mathbf{B}_{ij}$ be the crossproduct matrix between groups for $\mathbf{u}_i$ and $\mathbf{u}_j$ and $\mathbf{T}_{ij}$ be the corresponding matrix for the total group. We further define $\mathbf{B} = \{\mathbf{B}_{ij}\}$ and $\mathbf{T} = \{\mathbf{T}_{ij}\}$. The Q index can be $\mathbf{t}'\mathbf{B}\mathbf{t}/\mathbf{t}'\mathbf{T}\mathbf{t}$, which corresponds to formula 4. The K values of $\mathbf{t}$ may be restricted in different ways, analogous to the case of principle component analysis. To take a very restricted case, suppose that all x have k+1 categories and that we want to find a scale common to all x which gives maximal discrimination for the unweighted sum of the variables. Then $\mathbf{c}$ is fixed and $\mathbf{D}$ is constrained, so that there are only k-1 ranked values to determine. Other designs may also be treated.

In canonical correlation analysis we also use a second set of variables. Let y_i, i=1, ..., q, have m_i+1 categories, with scale vector $\mathbf{b}_i$ and $\mathbf{v}_i$

as its binary variables, such that $y_i = \mathbf{b}'_i\mathbf{v}_i$. Define $\mathbf{D}_b$ as a block diagonal matrix of order Mxq, having $\mathbf{b}_i$ on the principal diagonal ($M = \sum_1^q m_i$). We further need a weighted sum $\mathbf{d}'\mathbf{y} = \mathbf{d}'\mathbf{D}'_b\mathbf{v} = \mathbf{t}'_b\mathbf{v}$, where $\mathbf{v} = \{\mathbf{v}_i\}$. Finally, define the covariance matrices $\mathbf{S}_{vv} = \{\mathbf{S}_{v_iv_j}\}$, $\mathbf{S}_{uv} = \{\mathbf{S}_{u_iv_j}\}$ and $\mathbf{S}_{uu} = \{\mathbf{S}_{u_iu_j}\}$, of orders MxM, KxM and KxK, respectively. (For instance, the general element of $\mathbf{S}_{uv}$ is the covariance matrix between $\mathbf{u}_i$ and $\mathbf{v}_j$.) If we take Q as the squared correlation between $\mathbf{c}'\mathbf{x}$ and $\mathbf{d}'\mathbf{y}$ it can be written as $(\mathbf{t}'_a\mathbf{S}_{uv}\mathbf{t}_b)^2/(\mathbf{t}'_a\mathbf{S}_{uu}\mathbf{t}_a\mathbf{t}'_b\mathbf{S}_{vv}\mathbf{t}_b)$ and thus has the form according to formula 5. Of course, this is also applicable to multiple correlations, in which case q=1. As a new example of restrictions we can mention $\mathbf{c} = \mathbf{d}$ and $\mathbf{D}_a = \mathbf{D}_b$, provided that p = q and $k_i = m_i$. This is a reasonable constraint if $\mathbf{x}$ and $\mathbf{y}$ are the same variables, measured on two occasions.

For factor analysis, we are interested in scaling the manifest variables so that they fit, as well as possible, to a given factor model. Several goodness-of-fit criteria are conceivable, such as the common or generalized least squares criterion, a likelihood function or perhaps the index suggested by Tucker & Lewis (1973). In general, the factor model is not fully specified, meaning that there are factor parameters as well as scale values to determine. I imagine that this will imply an iterative process which 'walks' to and fro between scale values and parameters: starting with a set of scales, one estimates the parameters, which constitute the basis for a new set of scale values, and so on. If the fit is bad, the model is not compatible with data under any admissible transformations of the manifest variables, which is quite a general conclusion.

FINAL REMARKS

The intention of stability analysis is to get knowledge about how differently you can describe results due to different scales. We may imagine two classes of transformations: R(Q), for which the Q index of result is invariant, and R(C), the admissble class of scale transformations for a certain concept. The word 'admissible' has the following (loose) meaning: given a definition of a concept, the possible outcomes of the instrument chosen (to measure this concept) can be scaled according to any element in R(C) without fundamental objections as to a change of the concept. The most common example in educational research would be the class of linear transformations for R(Q) and all monotonic transformations for R(C).

Stability analysis is only necessary if $R(Q) \subset R(C)$, since otherwise the result is totally stable. (If $R(Q) \supset R(C)$, there are perhaps better. stable Q indices.) However, when $R(Q) \subset R(C)$ it is not unusual to choose a new Q, such that $R(Q) \supseteq R(C)$. Several devices for this can be found in nonparametric statistics. In my opinion, it is better to keep the original Q and sharpen the definition of the concept, such that $R(Q) \supseteq R(C)$ or, if this is not possible, to perform a stability analysis. Suppose we have a Q which we regard as a good description of data, but with $R(Q) \subset R(C)$. I cannot see any reason why we should lose information by choosing a new Q with $R(Q) \supseteq R(C)$ instead of performing a stability analysis.

It may be clear from the examples that one is sometimes most interested in obtaining an extreme value of Q, e.g. a minimal interaction or a maximal group differentiation. Discussion of such optimal scaling for more or less special cases is not rare in research literature. However, there may be occasions when one wants to report a typical Q value. This can be defined in several ways but let us take the expected value. This integral can be difficult to evaluate but type I simulations discussed earlier give information about the expected value. It is reasonable to use the arithmetic mean of the generated distribution.

Of the examples discussed above, such simulations have been performed for examples 1, 3 and 7 with 200 repetitions. One can, for instance, ask if Q from the scale with equally spaced values is typical. We answer by reporting the standardized Q value: example 1 gives 0.11, example 3 gives 0.14 for A, 0.13 for AB and 1.29 for between cells (the value of B is not defined due to superstability) and example 7 has values between 0.51 and 1.27. The answer is consequently not an unequivocal yes or no. Moreover, when a measurement is made on different populations and/or the data are treated with different methods, stability, minimal Q, typical Q or maximal Q may vary. In conformity with a test having different reliabilities for different situations, it can also have different scales for different situations, provided that $R(Q) \supseteq R(C)$.

For complex methods, it may be difficult to construct an effective algorithm for scanning R(C) in order to find special Q values. An alternative is to resort to selected transformations and investigate the variation of Q among these. I have done this for factor analysis, Larsson (1974), and found the results robust to (some) monotonic transformations. However, this is not a satisfactory approach and one must at least try to use more general methods, like the one proposed in this report.

Binary coding is not the only alternative here. It seems to me that one can also use polynomials. For an arbitrary, monotonic scoring, w, we represent x as a polynomial of w of degree k. Then $(w, w^2, \ldots, w^k)$ corresponds to $\mathbf{u}$ and the polynomial coefficients correspond to $\mathbf{a}$. But the formulation of the monotonic restriction is probably more complicated: instead of only ranking the elements of $\mathbf{a}$, you now have to rank weighted sums of the coefficients.

When k is large the use of a polynomial may be advantageous. For instance, a truly continuous variable implies $k = n-1$, an 'impossible' number of categories to work with. The problem is to reduce the number by putting together categories with lowest possible distortion of data. For a polynomial, the 'obvious' way is to reduce its degree but I do not know how to handle the binary variables.

Provided that the optimization routines turn out to be dependable, it is my intention to investigate the stability of some univariate statistical methods on various data sets. It may be interesting to know whether stability varies with e.g. different educational research areas, different statistical methods and different numbers of categories. The investigation will give access to programs designed to determine minimal and maximal Q (and perhaps typical Q) for some statistical methods. It seems to me to be more sensible to report minimal and maximal Q, perhaps along with Q for equally spaced scale values, than only the latter. Suppose that the latter Q is 0.25 in two different cases (possible range of Q: $0 \leq Q \leq 1$). Suppose further that $0.00 \leq Q \leq 0.75$ in the first case and $0.20 \leq Q \leq 0.30$ in the second case. I do not think that the stability information will cause one to judge the cases identically, although Q for equally spaced scale values is the same in both cases.

REFERENCES

Abelson, R.P. Scales derived by consideration of variance components in multi-way tables. I: Gulliksen, H. & Messick, S. (Eds.) Psychological scaling: theory and applications. New York: Wiley, 1960, 169-186.

Box, G.E.P. & Cox, D.R. An analysis of transformations. J. roy. statist. Soc. B, 1964, 26, 211-252.

Bradley, R.A., Katti, S.K. & Coons, I.J. Optimal scaling for ordered categories. Psychometrika, 1962, 27, 355-374.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Fhanér, S. Some comments in connection with Rozeboom's linear correlation theory. Psychometrika, 1966, 31, 267-269.

Jöreskog, K.G. Analyzing psychological data by structural analysis of covariance matrices. Research Report 73-2. Department of Statistics, University of Uppsala, 1973.

Kruskal, J.B. Analysis of factorial experiments by estimating monotone transformations of the data. J. roy. statist. Soc. B, 1965, 27, 251-263.

Kruskal, J.B. & Carmone, F.J. MONANOVA: A FORTRAN IV program for monotone analysis of variance (Non-metric analysis of factorial experiments). Behav. Sci., 1969, 14, 165-166.

Larsson, B. Obtaining maximal correlations by the construction of binary variables. Didakometry, No. 38, 1973.

Larsson, B. The influence of scale transformations: A study of factor analysis on simulated data. Didakometry, No. 40, 1974.

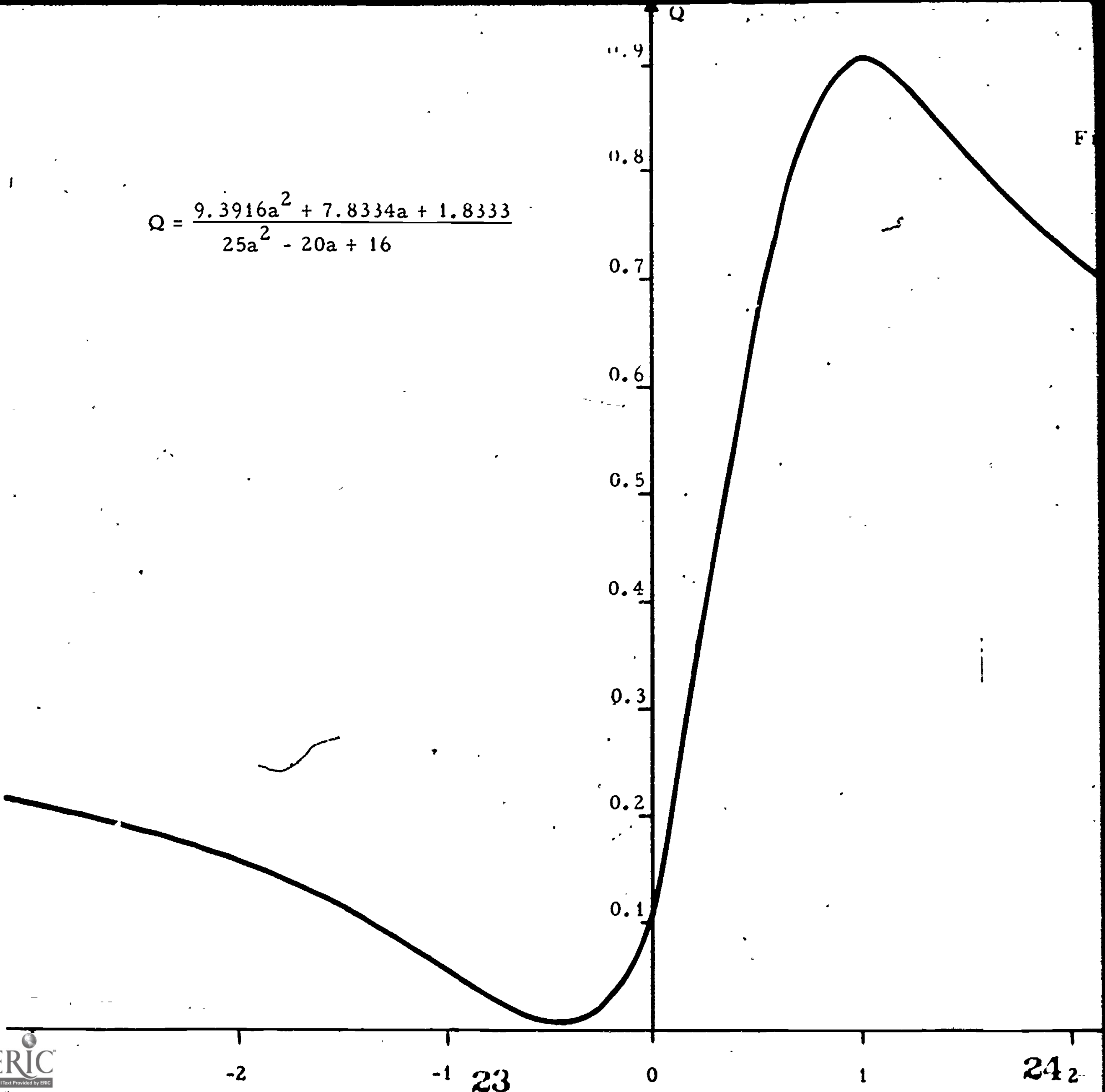
Lord, F.M. & Novick, M.R. Statistical theories of mental test scores. Reading, Mass.: Addision-Wesley, 1968.

Meredith, W.M., Fredriksen, C.H. & McLaughlin, D.H. Statistics and data analysis. Annual Rev. Psychol., 1974, 25, 453-505.

Tucker, L.R. & Lewis, C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 1973, 38, 1-10.

Tukey, J.W. Dyadic anova, an analysis of variance for vectors. Human Biology, 1950, 21, 65-110.

APPENDIX


$Q = \frac{9.3916a^2 + 7.8334a + 1.8333}{25a^2 - 20a + 16}$
Q
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-2
-1
0
1
2
F

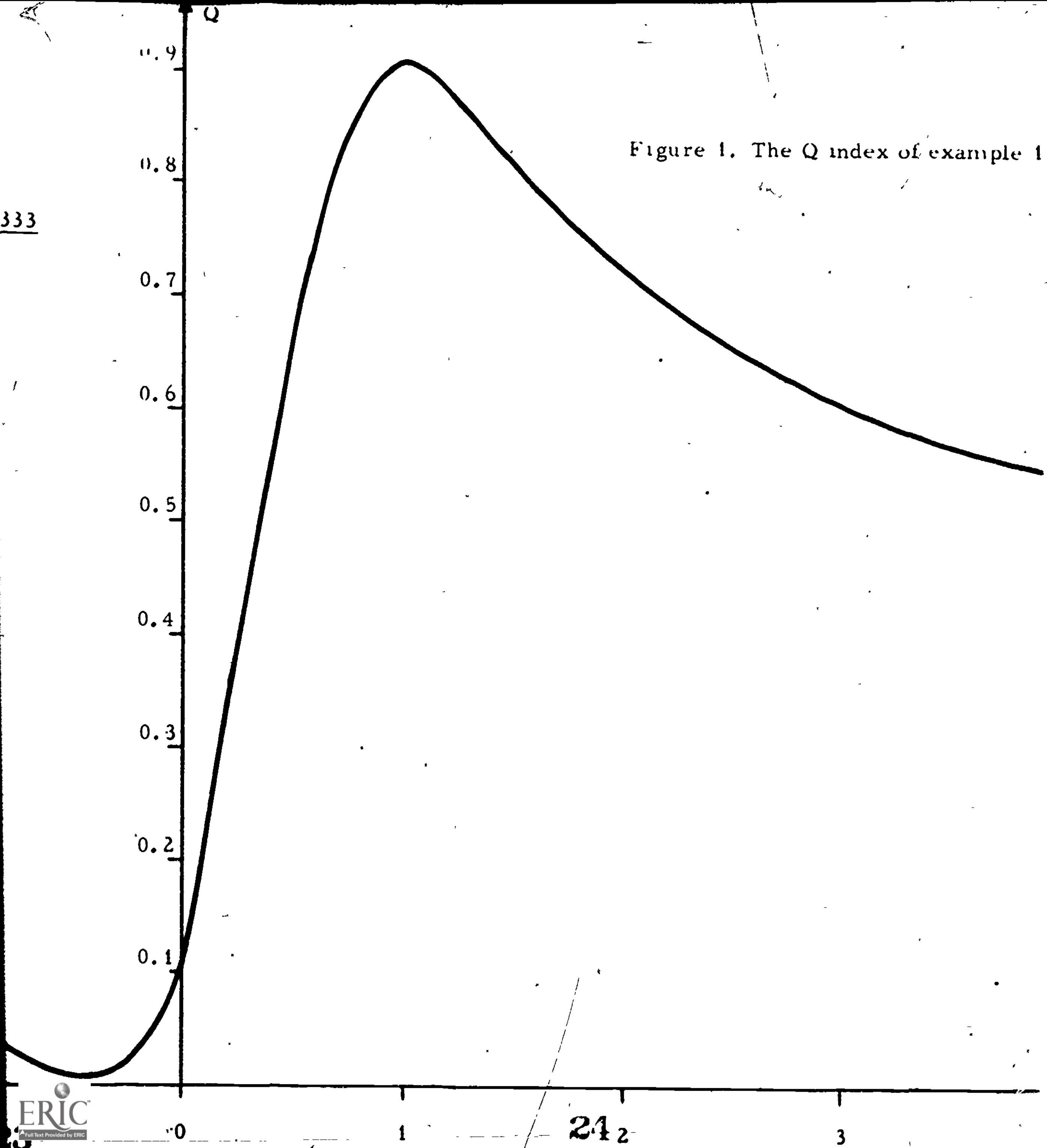
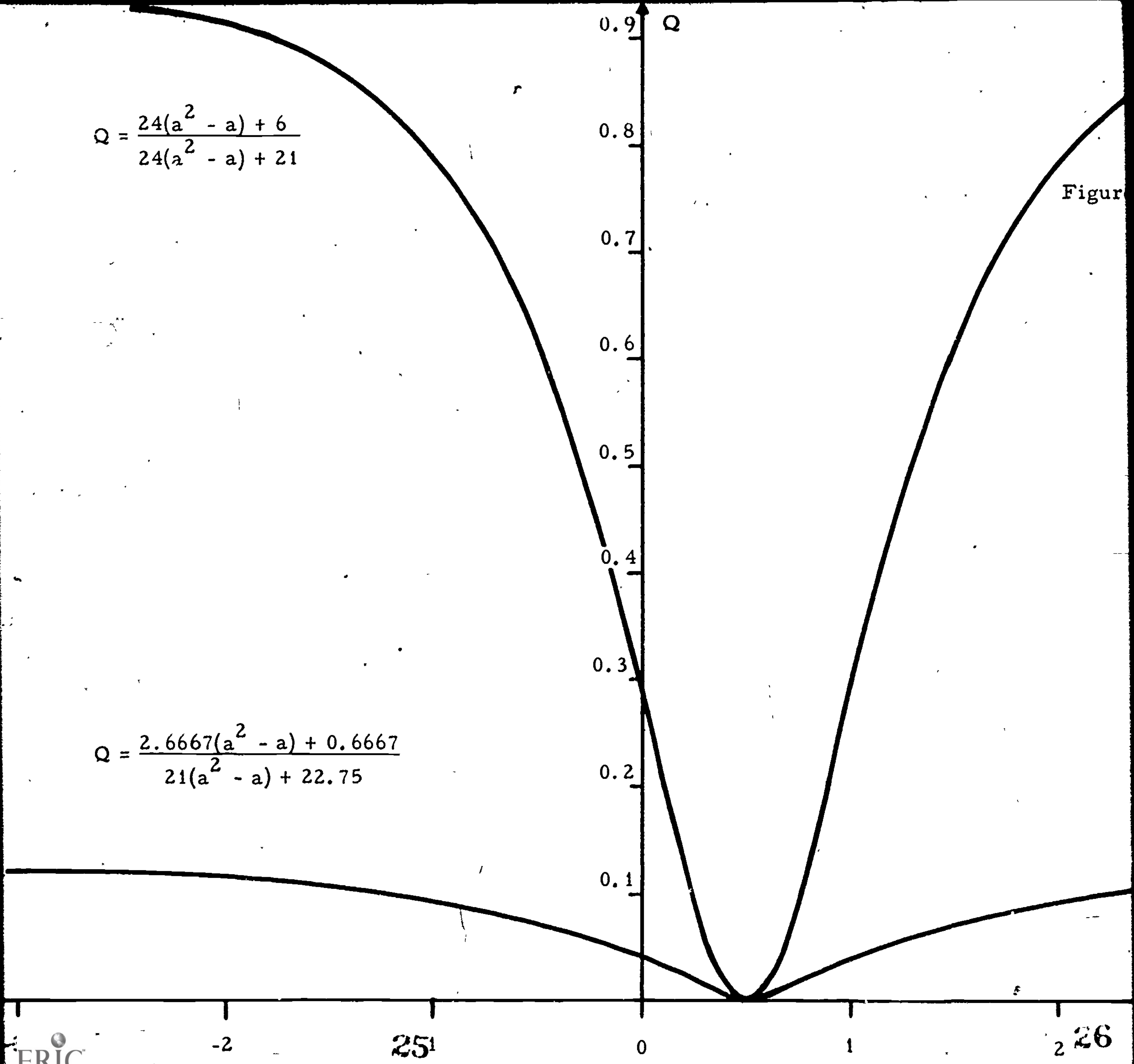


Figure 1. The Q index of example 1



Q
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
-2
0
1
2
$Q = \frac{24(a^2 - a) + 6}{24(a^2 - a) + 21}$
$Q = \frac{2.6667(a^2 - a) + 0.6667}{21(a^2 - a) + 22.75}$

Figur

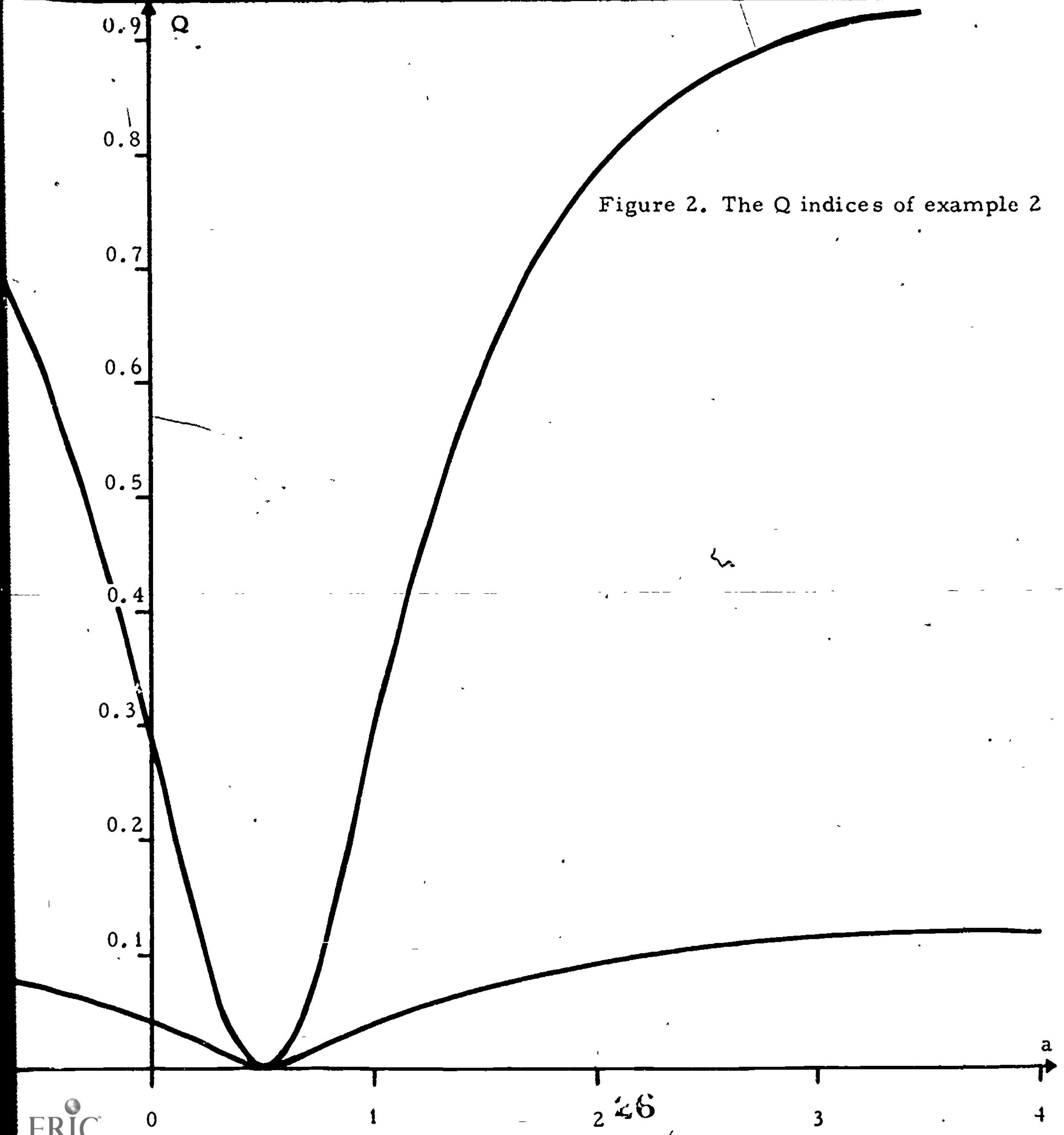
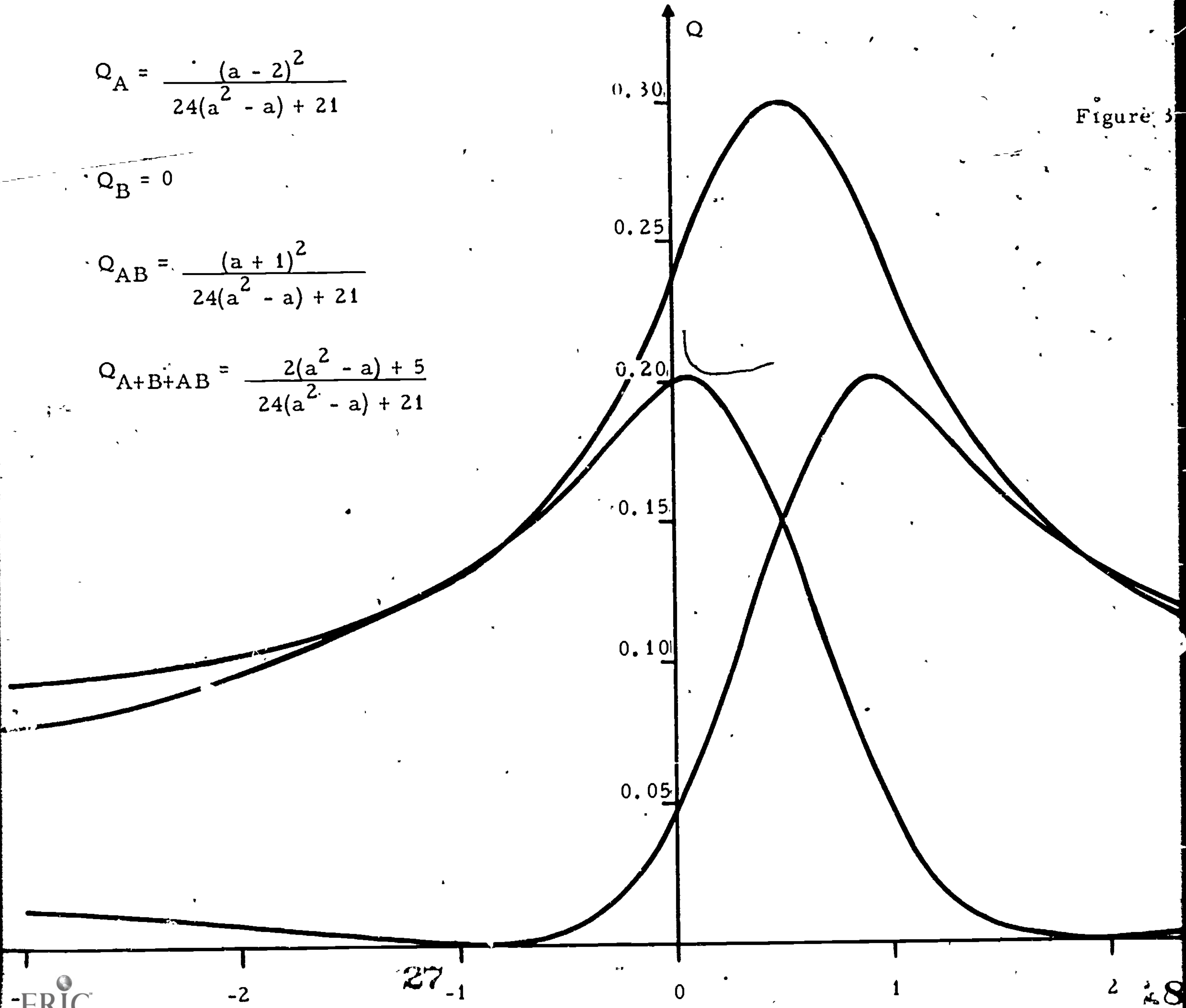


Figure 2. The Q indices of example 2

Figure 3



$$Q_A = \frac{(a - 2)^2}{24(a^2 - a) + 21}$$

$$Q_B = 0$$

$$Q_{AB} = \frac{(a + 1)^2}{24(a^2 - a) + 21}$$

$$Q_{A+B+AB} = \frac{2(a^2 - a) + 5}{24(a^2 - a) + 21}$$

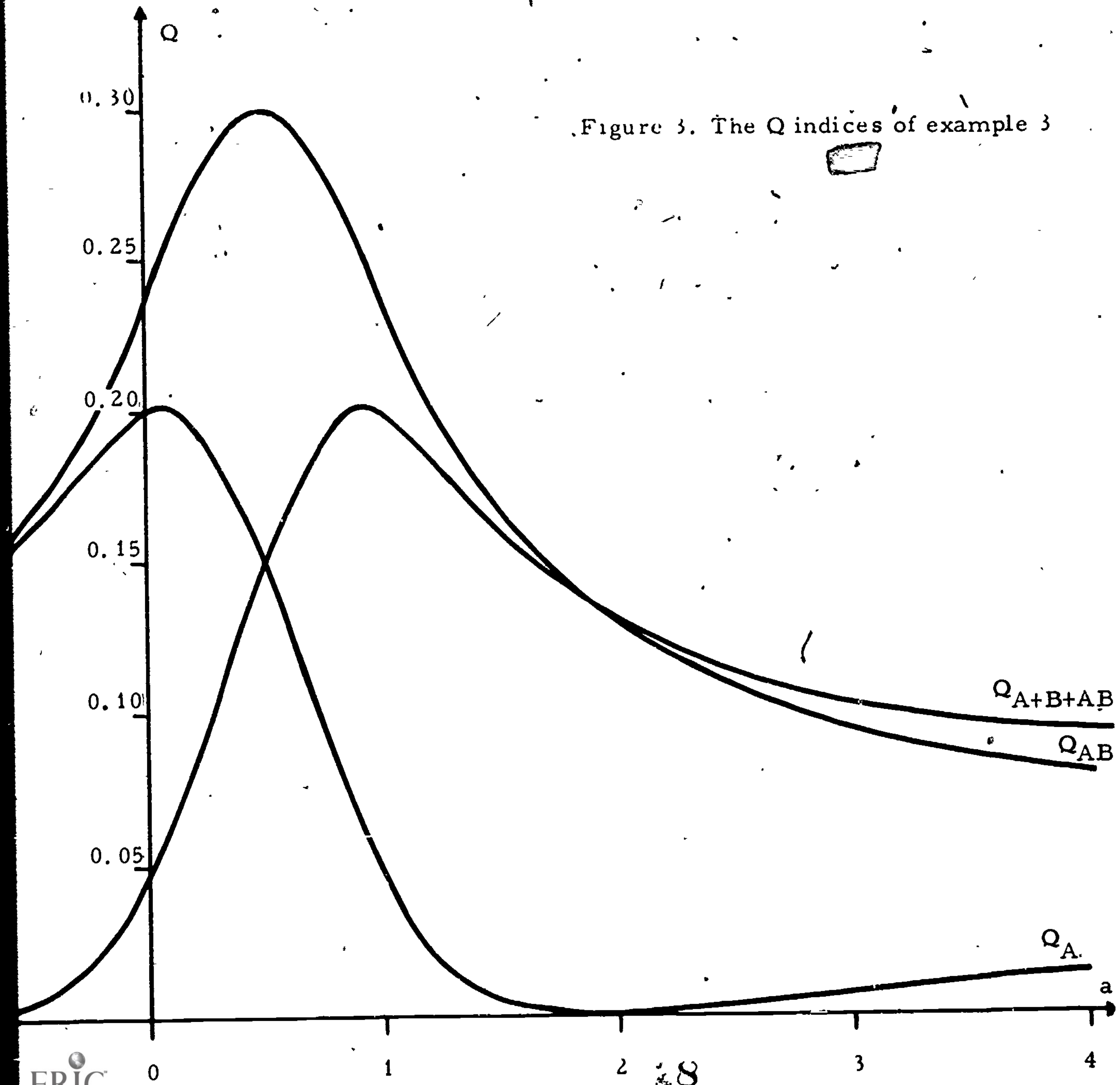
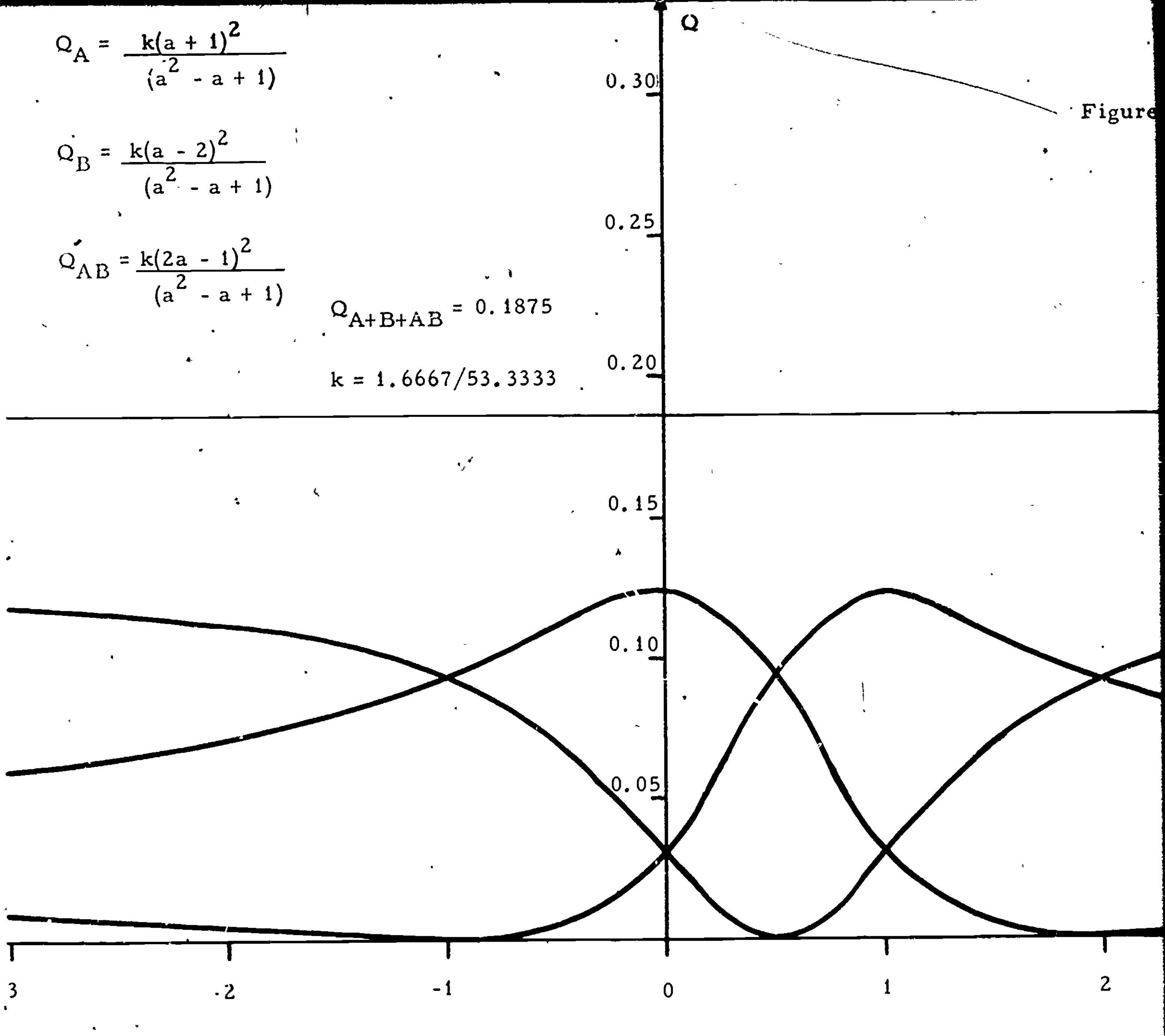


Figure 3. The Q indices of example 3


$Q_A = \frac{k(a+1)^2}{(a^2 - a + 1)}$
$Q_B = \frac{k(a-2)^2}{(a^2 - a + 1)}$
$Q_{AB} = \frac{k(2a-1)^2}{(a^2 - a + 1)}$
$Q_{A+B+AB} = 0.1875$
k = 1.6667/53.3333
Q
0.30
0.25
0.20
0.15
0.10
0.05
3
-2
-1
0
1
2

Figure

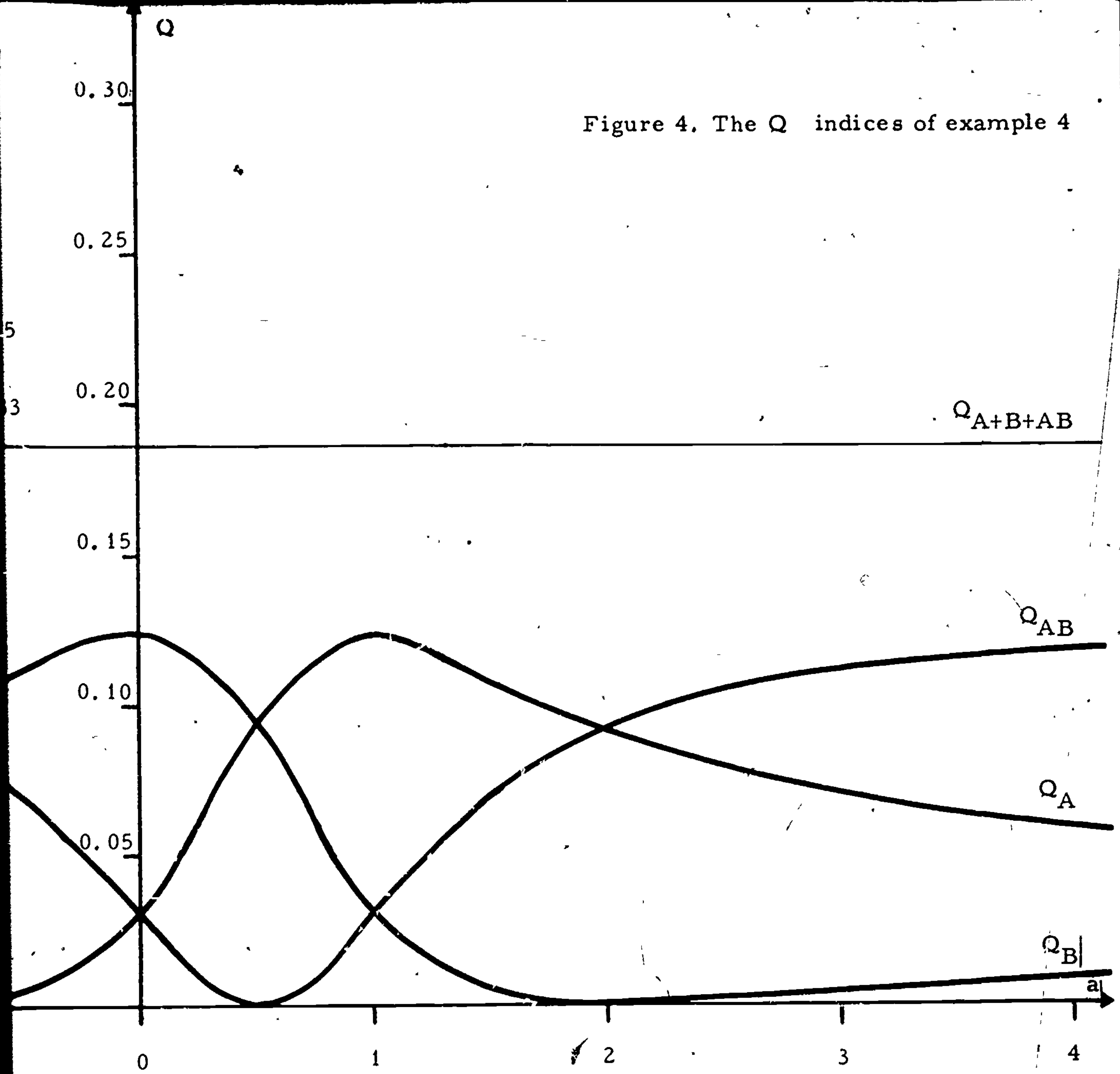
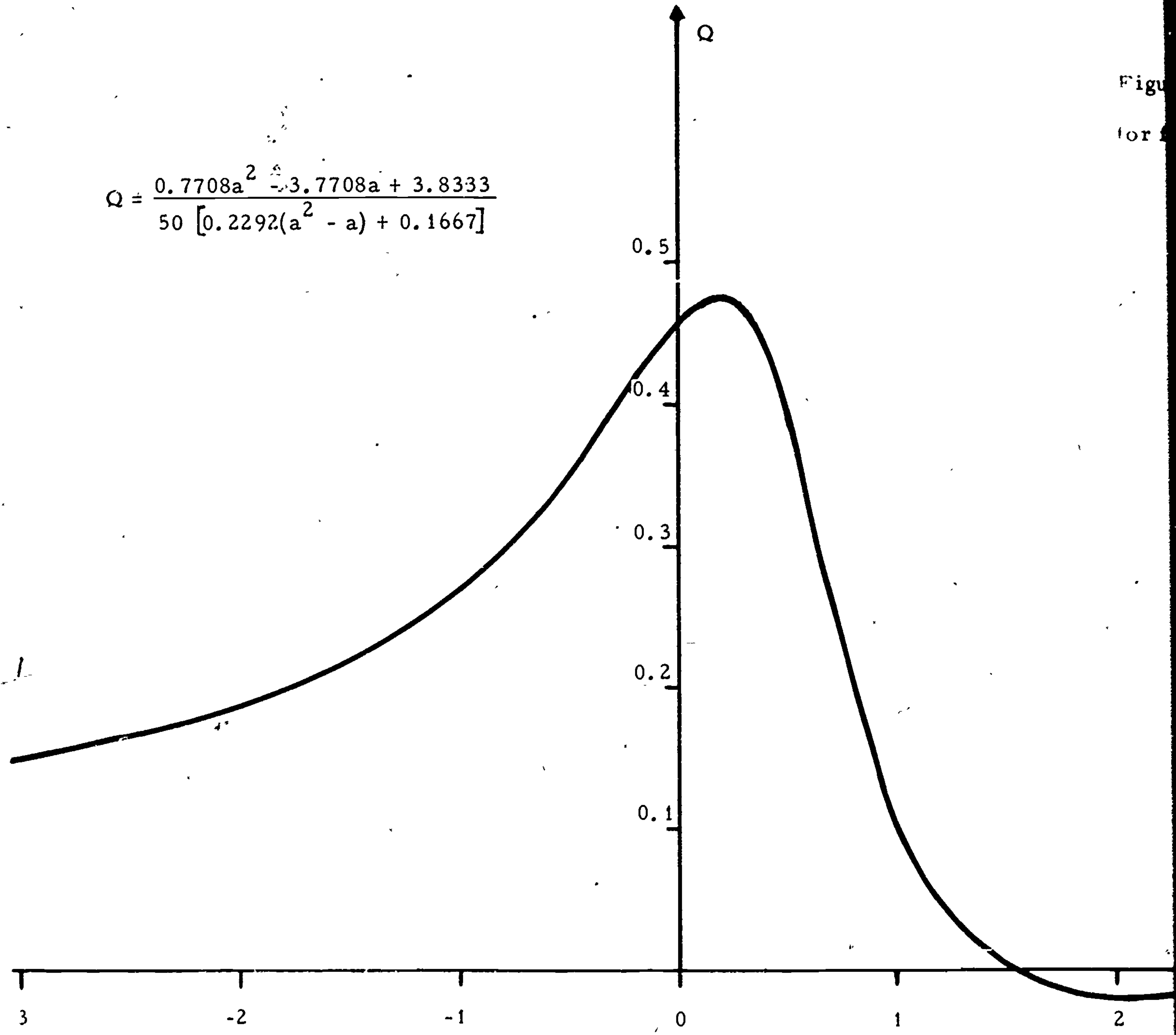


Figure 4. The Q indices of example 4



$$Q = \frac{0.7708a^2 - 3.7708a + 3.8333}{50\left[0.2292(a^2 - a) + 0.1667\right]}$$

Figu

for

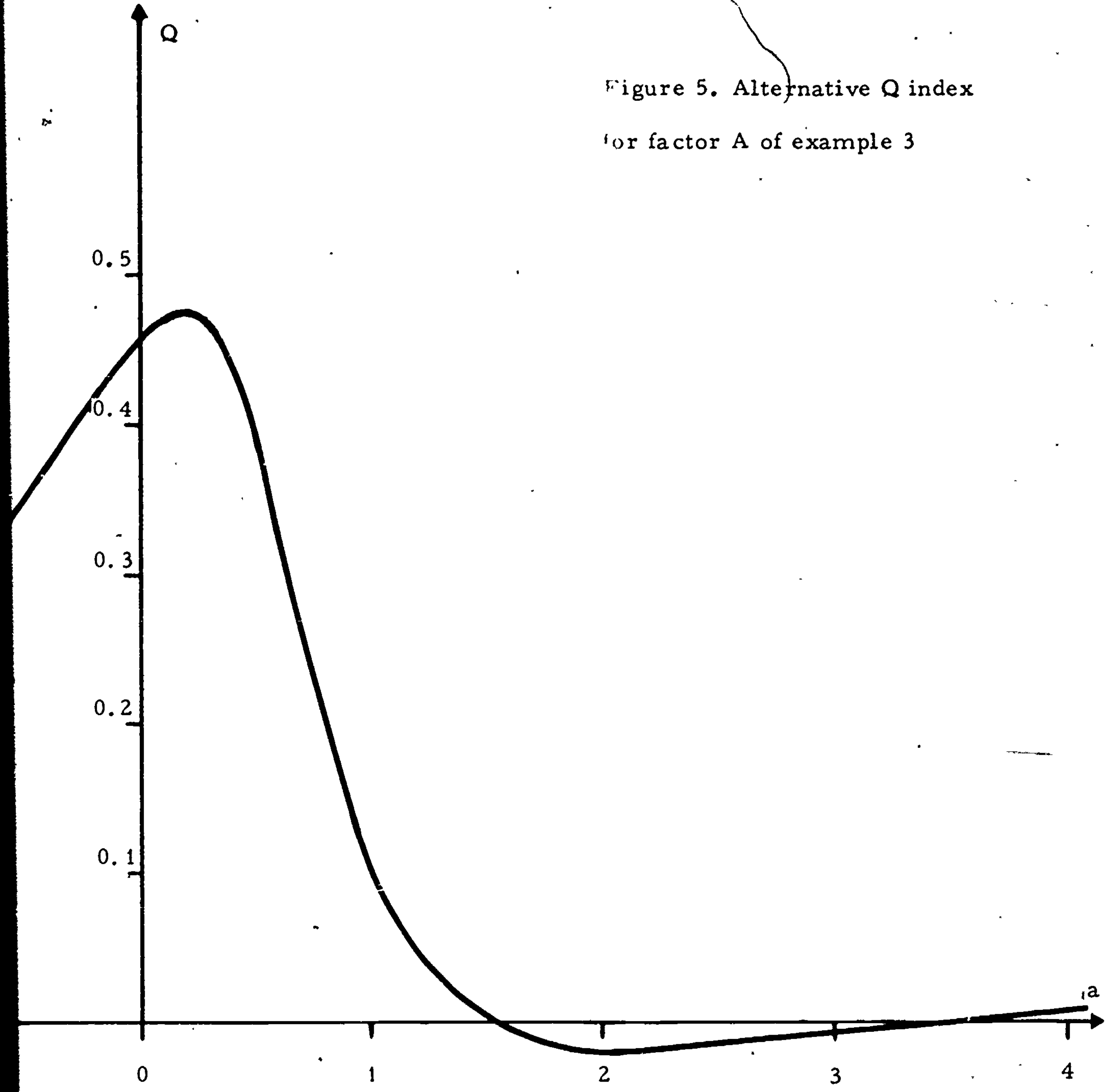
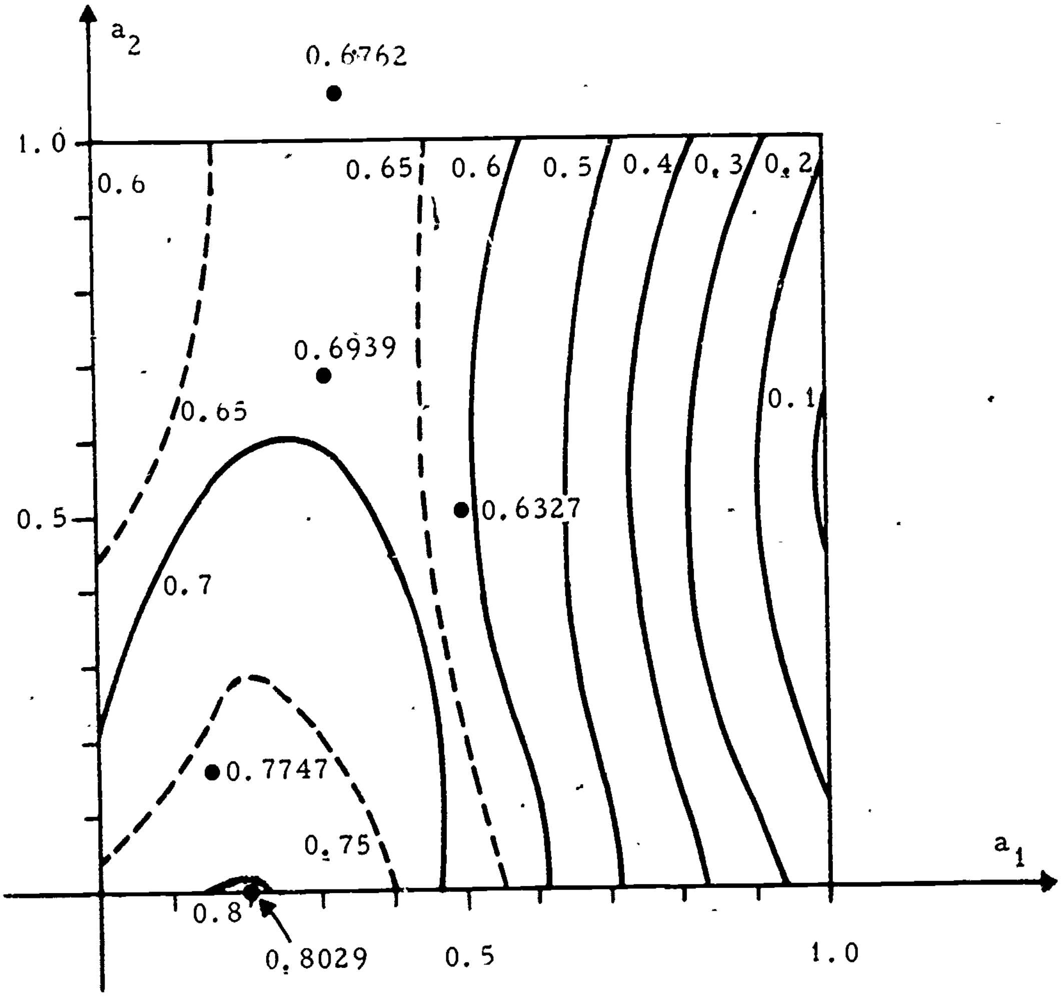


Figure 5. Alternative Q index for factor A of example 3



$$Q = \frac{-1.2a_1^2 + 2.8a_2^2 - 0.6a_1a_2 - 0.4a_1 - 2.6a_2 + 2.8}{3.7a_1^2 + 3.7a_2^2 - 2.4a_1a_2 - 2.6a_1 - 2.4a_2 + 3.7}$$

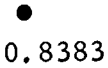
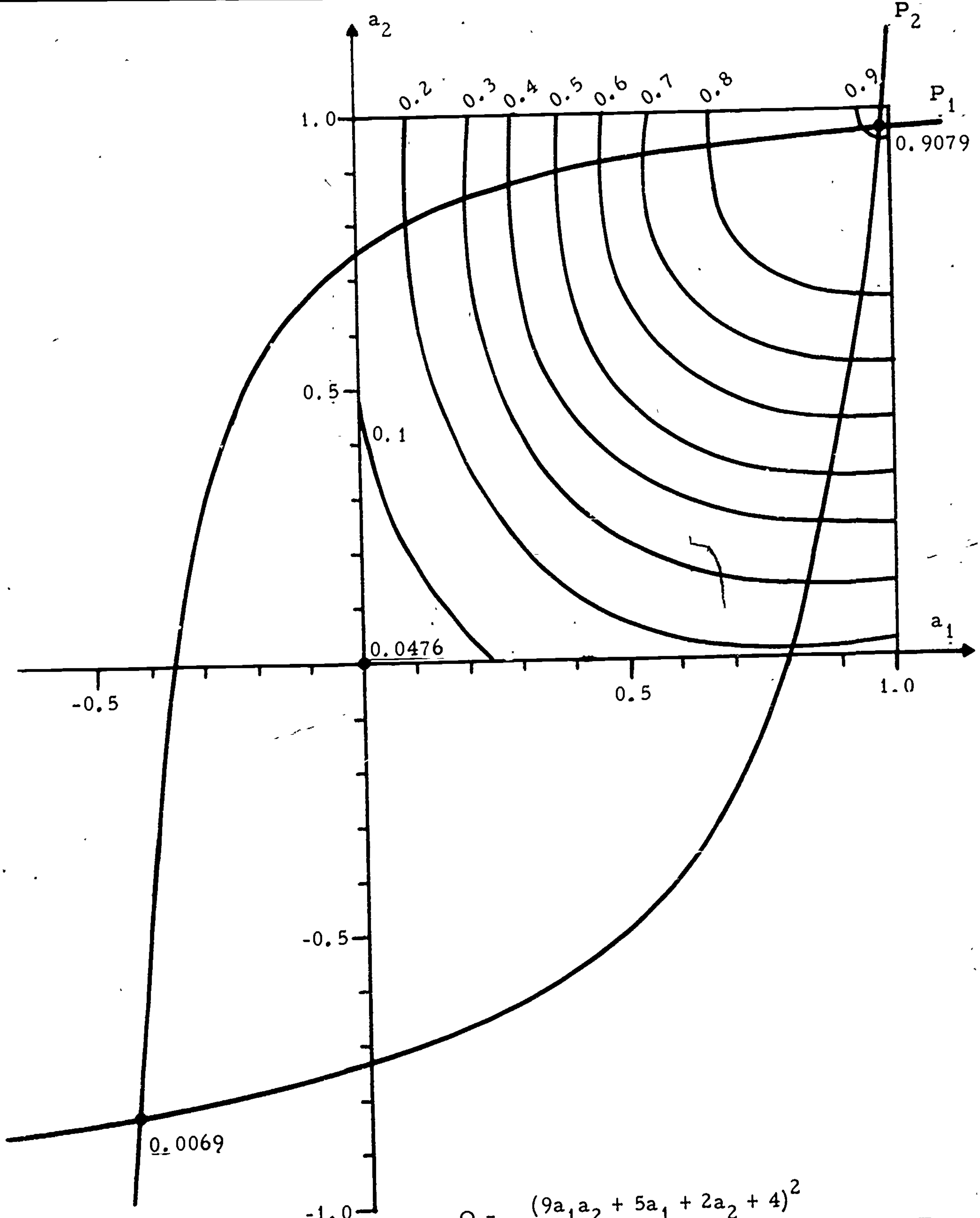


Figure 6. The Q index of example 5



$$Q = \frac{(9a_1a_2 + 5a_1 + 2a_2 + 4)^2}{\left[25a_1^2 - 20a_1 + 16\right]\left[24(a_2^2 - a_2) + 21\right]}$$

Figure 7. The Q index of example 6

Abstract card

Reference card

Larsson, B. The stability of results: Some examples of the effects of scale transformations. Didakometry (Malmö, Sweden: School of Education), No. 42, 1974.

This report gives some simple examples of stability for one factor and 2 x 2 factorial analysis of variance, reliability and correlations. The findings are very different: from superstability (no transformation whatsoever can change the result) to almost total instability. This is followed by a discussion of applications to multivariate analysis, and by some final remarks. It can be added that the technique can also be utilized for scaling variables to obtain a best fit to mathematical models other than those involved in usual statistical analysis.

Indexed:
1. Measurements
2. Transformations
3. Scales

Larsson, B. The stability of
effects of scale transform
Sweden: School of Educatio

Department of
Educational and
Psychological Research
School of Education
Malmö, Sweden

Reference card

xamples of the
etry (Malmö,
74.

of stability for
variance,
are very
mation whatso-
l instability.
tions to multi-
ks. It can be
ed for scaling
cal models
cal analysis.

Larsson, B. The stability of results: Some examples of the effects of scale transformations. Didakometry (Malmö, Sweden: School of Education), No. 42, 1974.