Appl. Statist. (1980),
29, *No.* 2, *pp.* 119–127

An Exploratory Technique for Investigating Large Quantities of Categorical Data

By G. V. Kass

University of the Witwatersrand, South Africa

[Received May 1977. Final revision December 1979]

Summary

The technique set out in the paper, CHAID, is an offshoot of AID (Automatic Interaction Detection) designed for a categorized dependent variable. Some important modifications which are relevant to standard AID include: built-in significance testing with the consequence of using the most significant predictor (rather than the most explanatory), multi-way splits (in contrast to binary) and a new type of predictor which is especially useful in handling missing information.

Keywords: AID; PREDICTION; CATEGORICAL DATA

1. Introduction

I consider the problem of parsimoniously describing a large data set. Consider many vectors each of whose elements is a categorized variable, one of which is the dependent variable. The remaining elements are predictors and their categories may or may not be orderable. A typical data set has several hundred vectors each with many elements.

Briefly, the technique, known as CHAID, partitions the data into mutually exclusive, exhaustive, subsets that best describe the dependent variable. The subsets are constructed by using small groups of predictors. The selected predictors may then be used in further analysis, prediction of the dependent variable, or in place of the total set in subsequent data collection.

When there are only a few predictors, or when the researcher has a preconceived description of the data, a log-linear model may be appropriate. This and other generalized linear models were described by Nelder and Wedderburn (1972). Everitt (1977) reviewed multidimensional contingency tables. The distinction between CHAID and these parametric methods is similar to that between standard AID and multiple regression.

1.1. *Comparison with* AID

The popular technique of automatic interaction detection (AID) was described by Morgan and Sonquist (1963a, b), Sonquist and Morgan (1964), Sonquist (1970) and Sonquist, Baker and Morgan (1971). It has been used in many fields; see Orr (1972) for an application.

In AID the data are successively bisected using one or other predictor, preserving the ordered nature of the categories within a predictor where appropriate. In the AID literature, such ordinal predictors are called monotonic, while purely nominal scaled predictors are called free. I shall introduce a new floating predictor that is most useful in practice.

AID operates on an interval scaled dependent variable and maximizes the between-group-sum-of-squares (essentially the *F*-statistic) at each bisection. In contrast, CHAID operates on a nominal scaled dependent variable and maximizes the significance of a chi-squared statistic at each partition, which need not be a bisection.

Standard AID is liable to misuse (see Doyle, 1973) and is open to the criticism that "... it has serious limitations because it never really takes into account the sampling variability inherent in the data" (Bishop, Fienberg and Holland, 1975, p. 360), a problem examined by Einhorn (1972).

CHAID tackles this problem by embedding the partitioning problem in a significance testing framework. This allows the formation and examination of multi-way splits which often leads to the conclusion that a predictor is indivisible according to the criterion. This should make it more suitable for inexperienced researchers and should counter doubts about AID and similar methods.

The selection procedure of AID favours predictors with more categories since the maximization criterion extends over more possibilities. A consequence of using significance testing in the decision-making process of CHAID is to nullify this bias.

1.2. *Comparision with* THAID (*Theta* AID)

The problem of a nominal dependent variable has been tackled before. The most pertinent reference is that of THAID discussed by Messenger and Mandell (1972) and described in detail by Morgan and Messenger (1973). The theta criterion for comparing binary splits is to maximize the sum of the number of observation in each modal category

When a dichotomous dependent variable is presented to AID and THAID they usually give different results. This has been verified by Morgan and Messenger in a computer simulation.

They support their use of the theta criterion mainly on intuitive grounds since they note (p. 18) : "Roughly speaking, AID tends to split giving outlying subgroups of small size much more than does THAID." The basis for this observation is not clear. If we assume that the phenomenon they observed arose for a monotonic predictor under the null hypothesis of a homogeneous group, then they are observing the result of Hawkins (1975) who showed that the position of the split point follows a U-shaped distribution. This is no indictment of the AID criterion since, under the null hypothesis, no split should be produced, and there is no reason to prefer a balanced split to an unbalanced one.

The standard AID (and hence χ^2) criterion has been subject to some theoretical advancement (Kass, 1975a; Scott and Knott, 1976) since the theta criterion was proposed. Knowledge of the theoretical behaviour of the theta criterion is lacking still.

2. Method of Analysis

Like AID, CHAID proceeds in steps, First the best partition for each predictor is found. Then the predictors are compared and the best one chosen. The data are subdivided according to this chosen predictor. Each of these subgroups are re-analysed independently, to produce further subdivisions for analysis.

The type of each predictor determines the permissible groupings of its categories, so as to build the contingency table with the highest significance level according to the chi-squared test. See, for example, Conover (1971, p. 149ff). This implies that there are enough observations to ensure the validity of this test. If this is not the case then some other criterion could be used, such as Fisher's exact test (Conover, 1971, p. 163ff), but I shall not pursue this possibility.

2.1. *The Algorithm*

Let the dependent variable have $d \geqslant 2$ categories, and a particular predictor under analysis $c \geqslant 2$ categories. A subproblem in the analysis is to reduce the given $c \times d$ contingency table to the most significant $j \times d$ table by combining (in an allowable manner) categories of the predictor. Conceptually, we may first calculate statistics $T_j^{(i)}$, the usual χ^2 statistics for the ith method of forming a $j \times d$ table ($j = 2, 3, \ldots, c$; the range of i depending on type of the predictor). Then, if $T_j^{(*)} = \max_i T_j^{(i)}$ is the χ^2 statistic for the best $j \times d$ table, choose the most significant $T_j^{(*)}$. The distribution of $T_j^{(*)}$ is discussed later.

In the case of a monotonic predictor or a dichotomous free predictor, the $T_j^{(*)}$ may be found by the procedure of Fisher (1958). This dynamic programming procedure is computationally of order c^2. When $d \geqslant 3$ a free predictor cannot have its categories ordered in general, and Fisher's method cannot be applied. Instead, the standard application of dynamic programming to permutation-type problems can be applied, as was done by Dreyfus and Law (1977, p. 69ff). This solution is computationally of order 2^c.

While dynamic programming is feasible on a computer for a single predictor at a single stage of the analysis, it would be unrealistic in a practical example where there are many predictors, and the analysis extends over a number of stages. I therefore propose an alternative method which, although it does not guarantee that the optimum solution will be found, has yielded very satisfactory results in practice. In contrast to dynamic programming, the computational effort is of order c for monotonic predictors, and of order c^2 for free predictors.

My proposal is to search for the $T_j^{(*)}$ in a stepwise manner. This procedure has parallels in other fields, for example multiple regression (Efroymson, 1965) and piecewise regression (McGee and Carleton, 1970). The full algorithm is as follows.

Step 1. For each predictor in turn, cross-tabulate the categories of the predictor with the categories of the dependent variable and do *steps 2 and 3*.

Step 2. Find the pair of categories of the predictor (only considering allowable pairs as determined by the type of the predictor) whose $2 \times d$ sub-table is least significantly different. If this significance does not reach a critical value, merge the two categories, consider this merger as a single compound category, and repeat this *step*.

Step 3. For each compound category consisting of three or more of the original categories, find the most significant binary split (constrained by the type of the predictor) into which the merger may be resolved. If the significance is beyond a critical value, implement the split and return to *step 2*.

Step 4. Calculate the significance (to be discussed later) of each optimally merged predictor, and isolate the most significant one. If this significance is greater than a criterion value, subdivide the data according to the (merged) categories of the chosen predictor.

Step 5. For each partition of the data that has not yet been analysed, return to *step 1*. This step may be modified by excluding from further analysis partitions with a small number of observations.

Steps 2 and 3 rely on the asymptotic partition of the total χ^2 statistic into a component, X, for the $2 \times d$ subtable that is being considered for collapsing, and a component for the resulting table after a merger. Kendall and Stuart (1961, Vol. 2, p. 577) claim that for finite samples "the asymptotic partition is good enough for most practical purposes". Kass (1975b) gave further details of the nature of the approximation. The component X is treated as a χ^2 variate with $d-1$ degrees of freedom, in the same way that F-tables are used in stepwise regression to help decide whether a variable should be included or excluded.

You may wonder if step 3 is really necessary. The algorithm precludes a split that would recreate a previous situation as long as the criterion for splitting a merger is stricter than the criterion for merging two categories. This latter condition is also necessary for the procedure to converge to a stable solution in a finite number of steps. In practice a merger is rarely split, but should be allowed to ensure near optimum results.

The algorithm is available as a PL/1 program to interested readers.

3. Significance of the Predictors

Step 4 of the algorithm requires a test of the significance of the reduced contingency table. If there has been no reduction of the original contingency table, a χ^2 test can be used. This test is conditional on the number of categories of the predictor, otherwise it must be viewed as conservative. The problems really arise when the original table has been reduced, because the algorithm has assured that it is the best possible for its size. If the table is reduced to two rows, then conditionally on a binary split, the results of Kass (1975a) may be applied to the case of a dichotomous dependent variable. This follows from the equivalence between χ^2_ν (in this case) and the rescaled F-distribution $\nu F_{\nu,\infty}$ (standard AID).

Noting the similarity between the Bonferroni critical values and those obtained by the exact theory, I wondered if the Bonferroni results could be used as a conservative first approximation in the case where the exact results were unknown. This idea has been partially validated by the computer simulation presented below. In effect, the procedure is to determine the number of ways a c category predictor *of a given type* can be reduced to r groups $(1 \leqslant r \leqslant c)$, and use it in the Bonferroni inequality to obtain a bound for the significance level. The formulae for calculating these multipliers for the three types of predictor allowed by CHAID are:

(1) *Monotonic predictors.* As in AID, a monotonic predictor is one whose categories lie on an ordinal scale. This implies that only *contiguous* categories may be grouped together. The Bonferroni multiplier is easily derived to be the binomial coefficient

$$B_{\text{monotonic}} = \binom{c-1}{r-1}. \tag{3.1}$$

(2) *Free predictors.* Again as in conventional AID, a free predictor is one whose categories are purely nominal. This implies that any grouping of categories is permissible. Finding the Bonferroni multiplier is a problem in partitions, or equivalently, by rephrasing as an appropriate occupancy problem, the multiplier can be derived (Feller, 1968, pp. 101–102) to be

$$B_{\text{free}} = \sum_{i=0}^{r-1} (-1)^i \frac{(r-i)^c}{i!\,(r-i)!}. \tag{3.2}$$

(3) *Floating predictors.* In many practical cases, the categories of a predictor lie on an ordinal scale with the exception of a single category that either does not belong with the rest, or whose position on the ordinal scale in unknown. I call it a "floating" category, and I call the predictor a floating predictor.

This situation typically arises when an investigation allows for an unknown or missing category. Missing information is often recoded by some technique such as that of Preece (1971) which allocates a possible score or value to the missing information. However, since such missing data do not upset AID-type analyses, it is usual (Morgan and Messenger, 1973) to assign a special category to them—the floating category.

Except for the floating category, grouping is only allowed for contiguous categories as for the monotonic predictors. The floating category, however, may stand alone or be combined with any other category or group of categories. The Bonferroni multiplier comes from a simple extension of the monotonic case:

$$B_{\text{float}} = \binom{c-2}{r-2} + r\binom{c-2}{r-1} = \frac{r-1+r(c-r)}{c-1} B_{\text{monotonic}}. \tag{3.3}$$

4. Empirical Investigation

A simulation under the null hypothesis was done to verify the efficacy of CHAID in practice. Specifically, one of the main aims was to control the type I error—the possibility of finding a spurious relationship in the data, indicated by a split.

Each simulation generated 480 data elements which were partitioned into 2, 3, 5, 8 and 10 category predictors. The simulation extended over 2, 3, 4 and 5 category dependent variables. Two hundred simulations were made for each type of predictor, making 10 trials for each combination of number of categories in the predictor and dependent variable.

The random data were obtained by generating 480 uniform random numbers u_i say, by a Tausworthe generator (see, for example, Whittlesey, 1968) and assigning the observation to category $[1 + du_i]$ for a d-category dependent variable. A c-category predictor was assigned cyclically $1, 2, \ldots, c, 1, 2, \ldots$ over the 480 data elements to ensure an equal allocation to each category.

In each case the merging criterion was significance at the 5 per cent level, and the splitting criterion was set to 4·9 per cent. The final result was only considered significant if the resulting contingency table was significant at the 5 per cent level after correction by its Bonferroni multiplier.

The results of the simulation are displayed in Table 1. Under the null hypothesis, each cell is expected to have 0·5 trials significant, making the total for each row 2·5 and the total for each column 2.

Each of the three sets of simulations had an approximately overall 5 per cent error rate matching the 5 per cent significance level that was used. The free predictor (b) had a 4 per cent error rate which is not significantly different from the expected 5 per cent. One may feel intuitively that the large Bonferroni multipliers for a free predictor could cause the test to be too stringent.

Table 1 is a satisfactory result and should instill confidence in CHAID users that they are not being misled by chance relationships.

TABLE 1
Number of significant results at 5 per cent level in data simulating the null hypothesis

(a) *Monotonic predictors*

	Number of categories in predictor					
Dependent variable	*2*	*3*	*5*	*8*	*10*	*Total*
2 categories	0	1	1	0	0	2
3 categories	1	0	0	1	0	2
4 categories	0	1	1	2	0	4
5 categories	1	0	1	0	1	3
Total	2	2	3	3	1	11/200 = 5½%

(b) *Free predictors*

	Number of categories in predictor					
Dependent variable	*2*	*3*	*5*	*8*	*10*	*Total*
2 categories	0	1	1	0	0	2
3 categories	1	0	0	1	0	2
4 categories	0	2	1	0	0	3
5 categories	1	0	0	0	0	1
Total	2	3	2	1	0	8/200 = 4%

(c) *Floating predictors*

	Number of categories in predictor					
Dependent variable	*2*	*3*	*5*	*8*	*10*	*Total*
2 categories	1	0	0	0	0	1
3 categories	1	0	1	1	1	4
4 categories	1	0	1	1	1	4
5 categories	0	0	1	0	0	1
Total	3	0	3	2	2	10/200 = 5%

5. Example

I have used CHAID successfully in several analyses. One of the larger ones run on an IBM S370/145 consisted of 954 observations. There were 7 monotonic predictors (with 2–9 categories), 11 free predictors (with 6–13 categories) and 26 floating predictors (with 10–14 categories). CHAID produced 13 final groups in just over 7 minutes.

For illustration I will now describe a smaller study of student data from the University of the Witwatersrand. Students are selected mainly on their matriculation examinations. There are several examination boards in South Africa, and there is a widespread belief that the standards set by the various boards are not equivalent. Various educationists have said that students who matriculate under some boards "do not perform as well as expected". This is partially borne out by a cross-tabulation of matriculation board and university performance, where students examined by a particular board consistently under-perform in various faculties. Rather than immediately blame the examination system, it was decided to investigate possible associations between the matriculation board and other background information about students.

The study covers 669 students for the Certificate in the Theory of Accountancy. The matriculation boards were classified into four categories: JMB (Joint Matriculation Board), NSC (National Senior Certificate), TUEC (Transvaal University Entrance Certificate) and Other. Seven predictors were included, the four of interest being:

(i) Year of entry (Free predictor).

(ii) Number of commercial subjects included in matriculation curriculum (Free predictor).

(iii) Type of school attended: conventional schools or tutorial colleges (called colloquially "cram colleges").

(iv) Matriculation mathematics symbol (this is a floating predictor with floating category labelled "?" for students who did not take mathematics).

Table 2 gives the details of the various stages reached by the algorithm in analysing the first predictor for the total group of students. Table 2a is the starting point with $T_4^{(1)} = 19{\cdot}1$ which, incidentally, is significant at the 0·024 level. To proceed to $T_3^{(\cdot)}$, all the six possible mergers of two rows are considered, the least significant being that between the first two rows which yields $X = 2{\cdot}2$ with 3 degrees of freedom in our previous notation. As X is clearly insignificant, the algorithm collapses the first two rows to yield Table 2b. Here there are three possible mergers, the least significant being that of the last two rows where $X = 3{\cdot}0$, again with 3 degrees of freedom. Finally, the algorithm considers Table 2c where the only possible merger has $X = 13{\cdot}7$ which, with 3 degrees of freedom, is significant at the 0·0035 level, and hence the merger is not performed.

As no splits of compound rows could be significant here, this final 2×4 table is accepted as "best", and its overall significance conservatively estimated using the Bonferroni multiplier of 7 ($c = 4$, $r = 2$ in equation (3.2)) as 0·024. Note that without the Bonferroni correction, a hasty investigator would consider the final result as significant beyond the 0·005 level.

After an examination in this way of each of the predictors in turn, the most significant was found to be the number of commercial subjects which partitioned the data into four groups as shown at the top of the dendrogram in Fig. 1. If you cast your eyes over this four-way split from left to right you will see the decrease in the proportion of students matriculating with the NSC board (labelled "N") as the number of commercial subjects decreases.

Each of the four subgroups are now candidates for further analysis. The leftmost two were not analysed as we set the program criteria to ignore subgroups of less than 100 students (a user option). The rightmost two groups (of 155 and 427 students respectively) were analysed separately in a similar manner to the orginal total group, and in each case the predictor "type of school" was found to be the best. An examination of these subgroups shows that it is evident that cram colleges have a predilection for the NSC board.

Two of the four subgroups at the second level of the dendrogram are eligible for further analysis. The group of 140 conventional school students who elected a single commercial subject was analysed, and the best predictor "English mark" did not meet the significance criterion set for this computer run. Hence, this group was not partitioned further, and the dendrogram terminates along this branch. Turn to the 412 conventional school students who eschewed commercial subjects; a significant difference was found using mathematics mark as a predictor. Fewer students seem to score well through the JMB (or those schools that fall under the JMB) than through the TUEC. The dendrogram is complete as no further partitions were significant.

TABLE 2

Details of the analysis of the predictor "Year of Entry" on the total group of students

(a) *Before any merging*

	Dependent variable				
Year	*Other*	*JMB*	*TUEC*	*NSC*	*Total*
1960	7	39	107	17	170
1961	13	34	112	17	176
1962	11	30	93	26	160
1963	13	34	80	36	163
Total	44	137	392	96	669

(b) *After one merger*

	Dependent variable				
Year	*Other*	*JMB*	*TUEC*	*NSC*	*Total*
1960, 1961	20	73	219	34	346
1962	11	30	93	26	160
1963	13	34	80	36	163
Total	44	137	392	96	669

(c) *After two mergers*

	Dependent variable				
Year	*Other*	*JMB*	*TUEC*	*NSC*	*Total*
1960, 1961	20	73	219	34	346
1962, 1963	24	64	173	62	323
Total	44	137	392	96	669

6. SUMMARY

Like its predecessor AID, CHAID is powerful technique for partitioning data into more homogeneous groups. While this type of analysis is often used as a precursor to a more parametric technique, it has been found frequently to be an end in itself. This is especially true in those fields where the typical method of analysis has been to produce and examine (if it is humanly possible) all cross-tabulations of the data. CHAID automates this to some extent by rejecting insignificant cross-tabulations, and immediately focusing the researcher's attention on the potentially useful subdivisions.

Some specific extensions have been made to AID. They include the floating predictor, and the notion of using the "most significant" split rather than the "most explanatory" which does not take into account the type of the predictor nor the number of its categories. This introduction of continual significance testing at each stage of the analysis has provided a criterion for assessing multi-way subdivisions of the data, which may yield a more effective analysis than the traditional binary splits that are often misleading and inefficient.

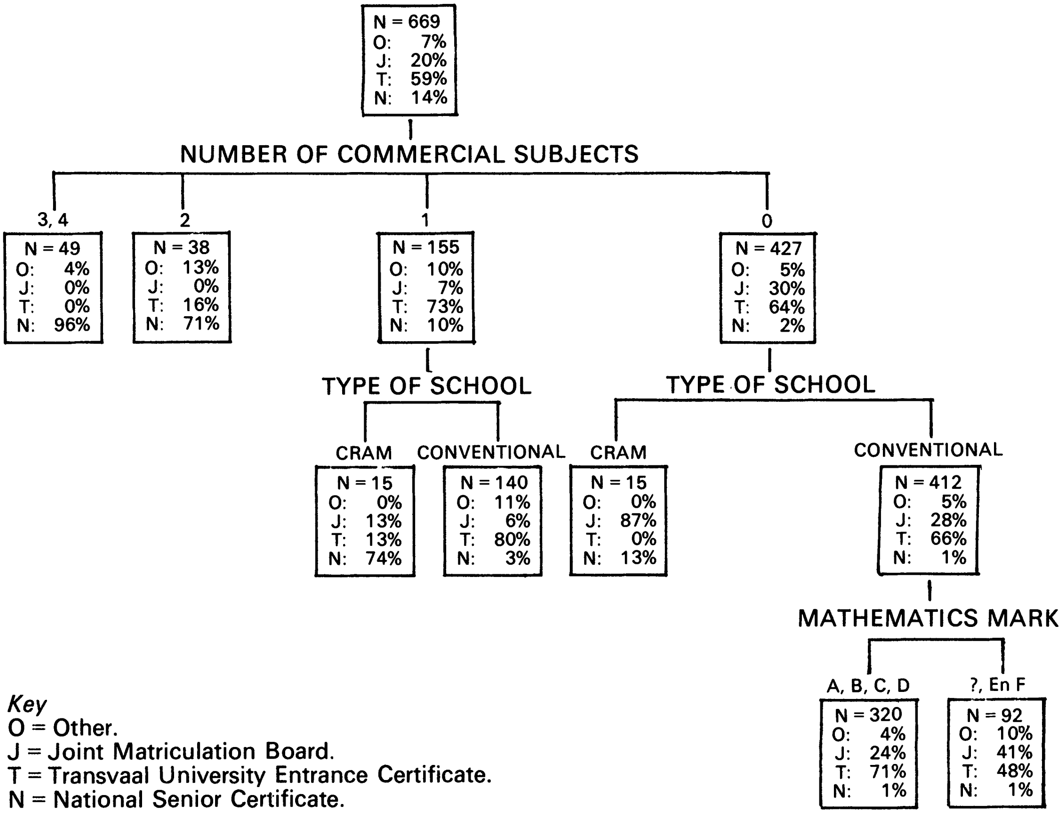

FIG. 1. Dendrogram of the analysis of South African matriculation boards.

7. Acknowledgements

Part of this work was done while the author was on long leave at Bar Ilan University, supported by a CSIR (S.A.) grant. The encouragement and advice of Professor D. M. Hawkins is gratefully acknowledged. Thanks are also due to the editor and referees for their helpful comments and detailed criticism of style.

References

Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis.* Cambridge, Mass. : MIT Press.

Conover, W. J. (1971). *Practical Nonparametric Statistics.* New York : Wiley.

Doyle, P. (1973). The use of the automatic interaction detector and similar search procedures. *Operat. Res. Quart.*, **24**, 465–467.

Dreyfus, S. E. and Law, A. M. (1977). *The Art and Theory of Dynamic Programming.* New York : Academic Press.

Efroymson, M. A. (1965). Multiple regression analysis. In *Mathematical Methods for Digital Computers* (A. Ralston and H. S. Wilf, eds). New York: Wiley.

EINHORN, H. J. (1972). Alchemy in the behavioural sciences. *Publ. Opin. Quart.*, **36**, 367–378.

EVERITT, B. S. (1977). *Analysis of Contingency Tables.* London : Chapman and Hall.

FELLER, W. (1968). *An Introduction to Probability Theory and its Applications.* New York : Wiley.

FISHER, W. D. (1958). On grouping for maximum homogeneity. *J. Amer. Statist. Ass.*, **53**, 789–798.

HAWKINS, D. M. (1977). Testing a sequence of observations for a shift in location. *J. Amer. Statist. Ass.*, **72**, 180–186.

KASS, G. V. (1975a). Significance testing in automatic interaction detec.ion (AID). *Appl. Statist.*, **24**, 178–189.

—— (1975b). Significance testing in, and some extensions of automatic interaction detection. Unpublished Ph.D. Thesis, University of Witwatersrand, South Africa.

KENDALL, M. G. and STUART, A. (1961). *The Advanced Theory of Statistics*, Vol. 2. London : Griffin.

MCGEE, V. E. and CARLETON, W. T. (1970). Piecewise regression. *J. Amer. Statist. Ass.*, **65**, 1109–1124.

MESSENGER, R. C. and MANDELL, L. M. (1972). A modal search technique for predictive nominal scale multivariate analysis. *J. Amer. Statist. Ass.*, **67**, 768–772.

MORGAN, J. N. and MESSENGER, R. C. (1973). THAID—a sequential analysis program for the analysis of nominal scale dependent variables. Survey Research Centre, Institute for Social Research, University of Michigan.

MORGAN, J. A. and SONQUIST, J. N. (1963a). Problems in the analysis of survey data : and a proposal. *J. Amer. Statist. Ass.*, **58**, 415–434.

——(1963b). Some results from a non-symmetrical branching process that looks for interaction effects. *Proc. of the Soc. Stats. Sec., ASA*, 40–53.

NELDER, J. A. and WEDDERBURN, R. W. M. (1972). General linear models. *J. R. Statist. Soc.* A, **135**, 370–384.

ORR, L. (1972). The dependence of transition proportions in the education system on observed social factors and school characteristics. *J. R. Statist. Soc.* A, **135**, 74–95.

PREECE, D. A. (1971). Iterative procedures for missing values in experiments. *Technometrics*, **13**, 743–753.

SCOTT, A. J. and KNOTT, M. (1976). An approximate test for use with AID, *Appl. Statist.*, **25**, 103–106.

SONQUIST, J. N. (1970). *Multivariate Model Building.* Michigan : Institute for Social Research, University of Michigan.

SONQUIST, J. N., BAKER, E. L. and MORGAN, J. A. (1971). *Searching for Structure* (*Alias*–AID–*III*). Michigan : Institute for Social Research, University of Michigan.

SONQUIST, J. N. and MORGAN, J. A. (1964). The Detection of Interaction Effects. Monograph No. 35. Survey Research Centre, Institute for Social Research, University of Michigan.

WHITTLESEY, J. R. B. (1968). A comparison of the correlational behaviour of random number generators for the IBM 360. *Commun. Ass. Comput. Mach.*, **11**, 641–644.