Fig. 5.1 Hypothetical Dendrogram of Education by Some Variables

```
         LEVEL OF EDUCATION
              HIGH %
              LOW  %
                   |
                   | INCOME
------------------------------------------
| HIGH/                                  | MEDIUM
| UNDISCLOSED                            | /LOW
|                                        |
% HIGH EDUCATION         HIGH EDUCATION %
% LOW EDUCATION           LOW EDUCATION %
|                                        |
(FURTHER SPLITS)                  (FURTHER SPLITS)
```

Since CHAID maximizes differences, one branch of the dendrogram will be identified with higher-than-average education and the other with lower-than-average education. It is then possible to follow "branches" of the tree in order to characterize groups. For example, suppose the (high, undisclosed) income people split on sex (male)/(female) and the (male) category of sex subsequently split on (urban)/(rural) we would then be able to characterize a group of, e.g., high-education people as being urban males with high or undisclosed incomes.

Ideally one would hope to find the same variable occurring at each branch on a given level of the tree. This would indicate a simple and stable model. *In actuality this seldom* occurs.

Within the CHAID program there are options for increasing or decreasing the level of Bonferroni significance at which variables are split. By definition this also affects the way categories within variables are grouped with respect to the predicted *variable. In the default mode*, CHAID will not carry out any analysis on on a group if the N of the group falls below fifty. This may be changed, though Hawkins and Kass (1982) recommend fifty as the minimum. Variables may also at the discretion of the user be included or excluded from the analysis. In our case a further possibility was provided: through the creation of two data sets we had the *option of analysing* data with missing values included or excluded.

Although CHAID was designed as a means of aiding the logistically daunting task of interpreting endless tabular runs, the quantity of output that may be produced using the options discussed above is still astounding. During analysis we found that through doing a large number of CHAID runs and producing rather laboriously for each run a complete dendrogram, a very good "*feel*" *of relationships within* the data was obtained. In particular, a few variables kept cropping up in the various trees, even if not always at the same levels, thereby indicating the "important" factors among the "relevant" ones.

What follows are a number of the more interesting dendrograms produced together with a brief discussion on each. In order that the reader may clearly follow our reasoning, some of the less relevant "early" runs are also included. In all cases the criteria for splitting was Bonferroni $p < 0.05$.

The actual variables finally used for CHAID as well as their specified categorizations are listed below. The initial variable set was larger than the final one, so that some variables represented in early dendrograms may not be on the following list.

Variables Finally Used in Chaid Analysis

In all cases "missing" values are labelled as "q".

PASS (q, pass, fail)

MATRIC RATING (q, various categories)

SEX (q, male, female)

RACE (q, Black, other)

MATRIC EXAM (q, Provincial/Internal, JMB, NSC, Indian Affairs)

PRE-MATRIC AGGREGATE, STD. NINE AGGREGATE, STD. EIGHT AGGREGATE, PRE-MATRIC MATHS, STD. NINE MATHS, PRE-MATRIC ENGLISH, STD. NINE ENGLISH (q, A, B, C, D, E and below)

NUMBER IN MATRIC CLASS (less than 23, 24-28, 29-32, 33-45,46 and more)

NUMBER OF SCHOOL CHANGES (q, none, one, two or more)

TEACHER AT ALL TIMES (q, yes, no)

REPETITION OF A STANDARD AT SCHOOL (q, never, one, more)

RELATIVE INDEX OF STUDY AT SCHOOL (q, harder, same, less)

MAIN INCOME EARNER (q,parents, other)

PARENTS DIVORCED/SEPARATED (q, yes, no)

PARENTS DECEASED (q, yes - one or both, no)

HOME LANGUAGE, FRIENDS LANGUAGE, SCHOOL LANGUAGE (q English, other)

NUMBER OF PEOPLE IN ROOM IN MATRIC (q, one/none, more than one)

FEELINGS ABOUT ATTENDING SCHOOL (q, positive, indifferent, negative)

EDUCATION OF FATHER, EDUCATION OF MOTHER (q, less than matric, matric, diploma, degree or better)

FINANCIAL HARDSHIP (q, yes, no)

REASONS FOR FINANCIAL HARDSHIP (q, familial, personal, environmental)

WHO FINANCES RESPONDENT (q, 1-self, 2-parents, 3-family 4-bursary, 5-bursary and self, 6-bursary and parents)

NUMBER OF NON-PRESCRIBED BOOKS READ PER MONTH (q, one, two, three, more)

PLACE OF RESIDENCE (q, home, university residence, other)

DIFFICULTY IN STUDYING AT RESIDENCE (q, yes, no)

TRAVELLING TIME (q, less than half an hour, half to one hour, one to one and a half, more)

MOTIVATION FOR DEGREE (q, interest, parental pressure, career)

FATHERS THOUGHTS DEGREE CHOICE, MOTHERS THOUGHTS (q, for, against)

FEELINGS ABOUT PLACE OF RESIDENCE (q, positive, negative, indifferent)

Our first (and rather naive) CHAID runs were centered on the concept of "who were passing how many subjects during the first year at university". The dendrogram (with five categories in the dependent variable) proved almost impossible to interpret. We accordingly then trichotomized pass/fail into clear pass - three or more subjects; close pass - two subjects; and fail.

The resulting dendrogram is reproduced below. All missing cases were excluded from the predicted variable (level of pass).

Fig. 5.2 Dendrogram of Level of Pass by All Variables

```
                LEVEL OF PASS
                    FAIL  27.3
                    PASS  35.3
              CLEAR PASS  37.4 (661)
                              |
                              |MATRIC AGGREGATE
      ----------------------------------------------
      |                                            |
      | A,B,Q,C                                    | D,E
    13.1                                         39.3
    31.4                                         28.2
    55.5 (374)                                   32.4 (287)
      |                                            |
      |  TIME WITHOUT TEACHER                      |WHY FINANCIAL
      |                                            |HARDSHIP
  -----------------------              ------------------
  |            |           |             |Q,PERSONAL        |FAMILIAL
  | 1          | 2         | Q,3,4       |ENVIRONMENTAL     |
11.8        54.5        11.9      35.8                   74.1
32.0        20.1        37.4      19.1                   15.7
56.2 (17)   25.4 (11)   50.8(346)45.1 (260)              10.2 (27)
  *           *            |         |                      |
     MAIN INCOME EARNER|         |  PRF MAT. MATH          $
     -------------------      -------
     |               |          |       |
     |Q,PARENTS      |OTHER     |Q      |ABCDE
   11.3          66.7        48.7    33.4
   29.5           0.0        31.1    27.4
   59.1 (343)  33.3 (33)     20.2(39) 39.1 (221)
     |            *            $        *
     |MATHS MATRIC
     ------------------
     |                |                  Key
     | Q,B,C,D        |
   2.9              15.5                 % FAIL
   5.9               5.6                 % PASS
   91.2  (272)      78.9 (71)            % CLEAR PASS   (N)
    $                 $                      |
                                             |SPLITTING VARIABLE
                                       ------------
                                       |
                                       |SIGNIFICANTLY DIFFERENT
                                       |CATEGORIES
                                       |
                                       *  END OF ANALYSIS
                                       $  FURTHER SPLITS NOT
                                          PRESENTED
                                       q  MISSING DATA
```

Immediately evident from the dendrogram is the considerable (but expected) importance of the respondents' school record. The matric aggregate of respondents was the main split (or predictor). Missing cases fell between a B and C aggregate.

At the lower end of the matric scale D and E aggregates are combined as significantly different from A,B and C aggregates. In comparison with the Bonferroni significance levels of all other variables entered into the analysis, matric aggregate is the most significantly related to the variable "good pass". All variables with a significant relationship to matric performance (see Chapter 3) were, however, also Bonferroni-significant.

If matric aggregate was excluded from the analysis, matric English, Afrikaans or mathematics would have been the first split variable. All had the same "grouping" - (A,B,q,C)/(D,E). The following variables were also Bonferroni-significant at the first split: repeating a standard at school - (q,more than one,none)/(one); financial hardship - (yes)/(q,no); reasons for financial hardship - (q)/(familial, personal, environmental).

Looking at the dendrogram it is the low-matric side that is of particular interest to us. The D and E aggregate people split on reported reasons for financial hardship. This predictor is a collapsed version of a subjective scale, and so must of course be interpreted with caution. During coding, meaning was imposed on responses in some way. Looking at the summary of results we found that financial hardship - (yes)/(no) - is also Bonferroni-significant. The strongest statement that can be made is that at the lower end of the matric scale students not experiencing some form of financial hardship pass at a higher rate than their fellow students.

It is gratifying to note that after only two splits, "high" achievers in the "low" matric categories (with a pass rate of 64.2%) are clearly differentiated from their fellow students within the "low" matric rating category (with a pass rate of 25.9%). Within the group "familial hardship", when it is split off from the rest of the "low" subgroup, the pass rate of the students is beginning to tend towards the overall pass rate of 72.7%!

The inverse form of reasoning may be applied to high-matric side of the dendrogram. If we follow the people who achieved a reasonable matric pass and then "under achieved" at university we arrive at a pass rate of 33.3% for the lowest "high" group - people who lacked a teacher for three months or longer in their last three years at school, and who also had someone other than a parent as the main income earner of the family. Looking at the summary of results for this group, reasons for degree choice - (q,interest,parental pressure)/(career) - would have been the split which would have occurred had the variable main income earner been suppressed.

Although we were, at this stage of the analysis, already finding the dendrogram output interesting, two factors prompted us to refine our approach. Although the predicted variable "good pass" is a more sensitive measure than pass/fail the purpose of the study was to look specifically at people in the "low" end of the matric rating scale. As the analysis proceeded we noticed that most of these people were in the middle category of "good pass". We also realised that in interpreting the dendrograms we were in fact adding pass to good pass in order to contrast pass with fail. We thus decided to collapse the predicted variable into pass/fail.

We were also particularly interested in the matric rating of respondents as this is what the University was using as a measure of tertiary ability. We had included the variable in a number of the CHAID runs categorised in a number of ways. These varied from the faculty exclusion categories then in force to completely arbitrary categories based on cell frequencies. We had expected the variable to constantly appear in the analysis at the level of the first split and it did not. We did know, however, from our own cross-tabulations and from the Senate report on first-year failure (University of the Witwatersrand, 1979), that the variable was strongly related to performance at first-year level. Deciding that we either had too many categories (up to nine) or that our arbitrary categorisations were simply inappropriate, we decided to see how matric rating, in as many categories as CHAID could admit, would be grouped by CHAID as a predictor of pass/fail. The categories of matric rating were as follows:

RANGE	%	N	RANGE	%	N	RANGE	%	N
0	9.0	65	25-26	5.8	37	35-36	7.1	47
1-15	0.3	2	27-28	10.0	66	37-38	5.5	36
16-20	1.5	10	29-30	6.0	40	39-40	3.9	26
21-22	3.5	23	31-32	7.7	51	41-45	12.1	80
23-24	6.7	44	33-34	8.6	57	GT 45	11.7	77

The following table shows how CHAID grouped these categories relative to the predicted variable pass/fail (missing cases were excluded from pass/fail).

Fig. 5.3 Dendrogram of Pass/Fail by Matric Rating

```
                PASS

            FAIL  24.5
            PASS  75.5 (661)
                    |
                    | MATRIC RATING
      ---------------------------------
      |             |             |
      |0-28         | 29-38       | GREATER 38
      |             |             |
    46.1          25.0           5.9
    53.8 (27.5%)  75.0 (39.3%)  94.1 (33.1%)
         |             |             |
         *             *             *
```

It might be argued that by this means we forced a particular structure on matric rating. This is unavoidably true. We were, however, confronted with a variable containing a large number of categories and we did not have prior information at our disposal apart from knowing that it was in some way related to pass/fail. The Faculty of Arts categorisations were themselves arbitrary. Matric aggregate was appearing as our best predictor, so matric rating obviously had the potential to be a more sensitive measure. Given the policy-use of matric rating which we were seeking to improve by our research, it made sense to combine the categories in terms of what we were attempting to predict. The benefit of the above procedure will become quite clear as we proceed.

Our first really useful (as opposed to merely informative) dendrogram was pass/fail with missing cases excluded against all variables in the data set, with matric rating combined in the above categories with respect to pass/fail. A summary of the dendrogram is produced in Fig. 5.4.

Fig. 5.4 Dendrogram of Pass/Fail by All Variables.

```
                          PASS
                     FAIL 24.5
                     PASS 75.5 (661)
                            |
                            | MATRIC RATING
                            |
------------------------------------------------------
|                          |                          |
| 0-28                     | 29-38                    | GT 38
|                          |                          |
46.2                      27.2                       7.5
53.8                      72.8                      95.2
(182)                     (213)                     (266)
 |                          |                          |
 | FIN                      | NMPM                     | NNLT
 |                          |                          |
  ----------            -----------           ----------
  | Q,2    | 1,3        | A,B    | D,E        | 1,2    | Q,4
  | 4      | 5          | Q,C    |            | 3      |
  |        |            |        |            |        |
52.8     22.5         11.1     37.1         28.1      4.7
47.2     77.5         88.9     62.9         71.9     95.3
(142)    (40)         (81)     (132)        (32)     (234)
  |        *            |        |            |        $
  | HARD   *            | DEC    | NMRT       | NFST   $
  |                     |        |            |
  -------           ---------   -------      -------
  |  Y   | Q,N      |Q,N    | Y | Q     | 29- |Q,N   | P,I
  |      |          |       |   |       | 38  |      |
72.2   46.6        6.8    57.1 68.0   29.9  71.4   16.0
27.8   53.8       93.2    42.9 32.0   70.1  28.6   84.0
(36)   (106)      (74)    (7)  (25)   (107) (7)    (25)
  |      |          |       |    |      |     |      |
  $      $          $       *    *      $     *      *
```

KEY

FIN - Who finances respondent
NMPM - Maths symbol in matric
NNLT - Length of time without a teacher
HARD - financial hardship
DEC - parents deceased
NMRT - matric rating
NFST - Feelings about staying at place of residence

(The categories of the response variables are shown earlier, on p. 128).

From the dendrogram it may be seen that matric rating combined in significantly different categories with respect to pass/fail becomes a better predictor of pass/fail than the other variables combined in the same respect. Caution should however be exercised in an attempt to apply these categories to following years. We found that the year of matriculation was significantly associated with matric rating (chi square = 17 for 4 d.f., $p < 0.01$). The categorisations of matric rating we have used should, however, be more applicable than any attempt to categorise matric rating exclusion levels arbitrarily. We had insufficient data on matric variables to attempt to find out why different years have different distributions of matric rating.

Among the "low" end of the matric rating scale a subgroup of 40 students with a pass rate of 77.5% was distinguished. This group was defined at the second split (immediately after the matric rating split) by the manner in which studies were financed. Those who were self-financed, partially dependent on a bursary scholarship or loan, or were being financed by family members comprised this group. A possible explanation is that respondents who exhibit some form of financial independence are generally older than their fellow students. Those who are financed by family members tend to have written the National Senior Certificate or the Indian Affairs Matric. The relevance of this split is that the pass rate of this group of people is higher than that of both second level splits of the matric rating 29-38 group. The pass rate of this group is also higher than the pass rate of the "low" split of the matric rating greater than 38 group. We have the very important situation where a group of 40 people have passed at a higher rate than 245 other respondents all of whom had significantly higher matric ratings than this group.

The situation is paralleled in the matric rating 29-38 group by students characterized by an A,B or C mathematics symbol who also have both parents living (pass rate equals 93.2%).

The above run was duplicated, with the variables matric rating and matric aggregate suppressed. The first split was on pre-matric aggregate - (A,B: pass = 96.6%)/(q,C: pass = 77.1%)/(D,E: pass = 51.8%). This split demonstrates both the pervasiveness of school record as a predictor of performance at first-year level and the importance of the pre-matric examination (discussed in the literature review). The second split for the (A,B) group occurred on "teacher at all times" - (yes)/(no). The "yes" group had a pass rate of 99.0% and subsequently split on "level of father's education" -

(matric,diploma)/(degree and better). This split demonstrates the already noted importance of a father's level of education in determining the success rate of individuals.

Unlike a simple cross-tabulation across all the cases concerned, CHAID enables one to "site" the relationships within subsets of cases. In the case of father's education it may be seen that the benefit goes to those who already have an excellent chance of passing first year (high matric marks). This ties directly into the "chicken and egg" problem of the literatur eview is education or disposable income the relevant predictor?

The (q,C) group subsequently split on financial hardship - (q,yes)/(no). The pass rate of those enduring some form of financial hardship had dropped to 20.1%. If the variable financial hardship was suppressed, reasons for financial hardship became the split - (q, personal, environmental)/(familial). The "low" (D,E) end of the dendrogram split on "reasons for financial hardship" - (q, personal, environmental)/(familial). The result seems to lend credibility to our somewhat brutal re-categorisations of the variable "reasons for financial hardship".

The pass/fail dendrogram was also re-run with missing categories included in the pass/fail variable. This had the unusual result of causing the q category of the matric rating variable to appear as a unique category. Following this matric rating (missing) branch of the dendrogram, we found that the second split was on feelings about school attendance. Once again the (q) category was split from the other categories, with the missing group on feelings about school attendance having pass = 14.3%, fail = 14.3%, and missing = 71.4% whereas all other categories on feelings about school had pass = 67.7%, fail = 30.6%, and missing = 1.6%. Only 3.6% of the entire sample had missing values for pass/fail, yet the proportion of missing values for pass/fail within the missing group on feelings about school attendance is 71.4%. However, the N of this group was only 14. An attempt was made to trace these people and add their responses to the data. It was not, however, very successful. We would speculate that most were foreign qualifications (not included as valid in the matric rating categories).

We generally found that including missing cases in the predicted variable pass/fail in turn gave rise to splits on missing cases. Except for the case of pass/fail this tends to identify groups of people with no interest in completing questionnaires fully, rather than producing meaningful results.

Following up the pattern observed in the general cross-tabulations, we performed a number of CHAID runs with matric rating (in significant categories with regard to pass/fail) as the predicted variable. This was done in an attempt to clarify the observation that some variables affected only matric rating or pass/fail, while other variables affected both. The resulting dendrogram is reproduced below. We decided to include missing values in this run since they formed a relatively large (10%) proportion of the matric rating variable.

Fig. 5.5 Dendrogram of Matric Rating by all Variables.

```
                              MATRIC RATING
                          q          9.8
                          0-28      17.5
                          29-38     33.7
                          GT 38     36.9
                                   (661)
                                     |
                                     | NPMAG
 ------------------------------------------------------------
 |                    |                    |                  |
 | q                  | A,B                | C                |D,E
 |                    |                    |                  |
30.8                 7.6                  4.5                9.5
24.4                 2.5                 10.2               31.4
21.8                 6.8                 41.6               49.5
23.1                83.1                 43.7                9.5
(78)               (118)                (245)              (220)
  $                                        |                  |
  $              NMEX          PME         |         YRM      |
                    |                      |                  |
        ----------------    ----------------  ----------------
        |              |    |              |  |               |
        | q        P,C |    | q,A,B,C   D,E|  |q              |
        |              |    |              |  |79,80    81,82 |
      54.5           2.8  3.6            7.8 18.2           4.9
       0.0           2.8  5.7           27.5 35.1          29.4
       9.1           6.5 39.2           51.0 33.8          58.0
      36.4          87.7 51.5           13.7 13.0           7.7
      (11)         (107) (194)          (51) (77)          (143)
        *             $    $              $    $              $
        *             $    $              $    $              $
```

KEY
NPMAG - Pre-matric aggregate
NMEX - Matric exam written
PME - Pre-matric English
YRM - Year of matriculation

At first glance, the dendrogram presented is not as enlightening as we hoped it would be. It serves mainly to confirm what we already knew, viz., that students performed with relative consistency through their school years and that subsequently school record was generally the best predictor of matric performance. Demonstrated above is, once again, the tendency for missing data to split on missing data. At the level of the first split, if all school record variables were suppressed, "which matric exam written" in the categories - (National Senior Certificate,Indian Affairs)/(Provincial)/(q) - appears as the first split.

However, the real interest of Fig. 5.5 is in the variables that do not come up! For when we compare our "pass/fail" dendrogram in Fig. 5.3, to our "matric rating" dendrogram in Fig. 5.5, we found that apart from school-record variables there seemed to be less overlap in the predicting variables than might have been expected. Since matric rating is the best predictor of pass/fail, one would have ideally expected that the non-school variables which predict matric rating would also be predictors of pass/fail. That this was not always the case confirmed our initial impression from the cross-tabulations.

This is an interesting issue, with policy importance. If there are variables within the one set of predictors which do not occur in the other, the implication is that there is something revealing which is not being taken into account by the body evaluating the potential performance of the student at university. For example, assume a variable is actually a significant predictor of matric rating, but does not appear as significant in the list of potential pass/fail predictors. Two cases are possible. If the variable negatively influences matric rating, the respondent will do better at university than matric marks indicate he or she will. If the variable has a positive effect on matric rating, then the student will not do as well at university as expected. The situation is likewise applicable to variables predicting pass/fail while not predicting matric rating.

A problem does, however, immediately arise. Matric rating is in itself an important predictor of pass/fail. How are we to know that a variable, while not "significantly" a predictor of pass/fail, has nevertheless influenced matric rating to such an extent that the variable does actually (though not directly) influence pass/fail? In order to clarify this situation we needed to apply log-linear analysis to the data set.

Our input into log-linear analysis was all variables (for both pass/fail and matric rating) which would, in the absence of a variable that was more Bonferroni-significant, have been the first "split" in terms of the predicted variable. These variables had all been grouped by CHAID into significantly different categories with regard to the predicted variable. It can be seen now that this ability of

CHAID to re-categorise variables in terms of their significant association with chosen dependent variables, was the main thrust of its use in our context.

. LOG-LINEAR ANALYSIS

As noted in the previous chapter, we need to turn to log-linear analysis in order to resolve which background variables affect mat scores without affecting pass/fail or affect pass\fail directly without affecting matric scores, or affect pass\fail indirectly via their affect on matric scores. More tech y put, we are checking on the presence or absence of mat rating - because of the way it is relied on administratively as an intervening variable between various antecedent factors and pass/fail at the end of the first year of university.

Presented below is a very brief and simplistic exposition of how log-linear analysis actually works. The overall thrust of the method is to analyse contingency tables multiplicatively, i.e., in terms of ratios among frequencies, thereby emphasizing the "pull" of the association between the variables rather than the respective levels of the quantities being measured.

Three concepts are central to an understanding of the log-linear method: odds ratios, effects, and models.

To understand the basis of odds ratios, consider the following hypothetical tables.

Table 6.1 Hypothetical Tables to Demonstrate Log-Linear Analysis

		HAIR COLOUR (A) Blonde	Brunette	Total
EYE	Blue	a=20	c=30	50
COLOUR	Brown	b=30	d=45	75
	Total	50	75	125

		HAIR COLOUR (B) Blonde	Brunette	Total
EYE	Blue	a=30	c=20	50
COLOUR	Brown	b=20	d=55	75
	Total	50	75	125

We define "conditional ratios" (Davis, 1978) as a/b and c/d. (Of course a/c and b/d would work as well.) In table A, both take the value of 2/3, telling us that in that (rather unlikely) sample, the incidence of blue as against brown eyes is the same amongst blondes as amongst brunettes: in other words, eye colour is independent of hair colour. (The table also tells us that blondes are less numerous than brunettes, in which the method is less interested.) The "odds ratio" for the table is defined as (a/b)/(c/d), in this case being one. So an odds ratio of unity represents independence between the two variables.

But now look at table B above. The respective incidences of blondes versus brunettes and blue versus brown eyes has been preserved. But the conditional ratio of blue to brown eyes for blondes is a/b = 30/20 = 1.5, and for brunettes it is c/d = 20/55 = 0.36. In other words, hair and eye colour are not independent in this sample: blue eyes tend to go with blond hair, brown eyes with brunette; and the odds ratio for the table, 1.5/0.36 = 2.75, is a measure of the "pull", i.e., the strength of the association.

The terminology becomes intricate with the addition of extra variables. But the principle is easy to understand. If one took table A as referring to Germany and table B to Sweden, we have an "interaction": the strength of the association between hair colour and eye colour is less in the former than the latter milieu, which could be summarised by noting that the ratio of the odds ratios is 1/2.75, i.e., less than one!

These more or less complex odds ratios are used to quantify what are called "effects". An effect amounts to variations among cell frequencies which reflects an association between a pair of variables, and in the case of more than two variables, an interaction. According to Davis "The analyst

But now look at table B above. The respective incidences of blondes versus brunettes and blue versus brown eyes has been preserved. But the conditional ratio of blue to brown eyes for blondes is $a/b = 30/20 = 1.5$, and for brunettes it is $c/d = 20/55 = 0.36$. In other words, hair and eye colour are not independent in this sample: blue eyes tend to go with blond hair, brown eyes with brunette; and the odds ratio for the table, $1.5/0.36 = 2.75$, is a measure of the "pull", i.e., the strength of the association.

The terminology becomes intricate with the addition of extra variables. But the principle is easy to understand. If one took table A as referring to Germany and table B to Sweden, we have an "interaction": the strength of the association between hair colour and eye colour is less in the former than the latter milieu, which could be summarised by noting that the ratio of the odds ratios is 1/2.75, i.e., less than one!

These more or less complex odds ratios are used to quantify what are called "effects". An effect amounts to variations among cell frequencies which reflects an association between a pair of variables, and in the case of more than two variables, an interaction. According to Davis "The analyst working within Goodman's system views the cell frequencies in his cross-tabulation as a dependent variable" (Davis, 1978:242). These cell frequencies (effects stated in terms of odds ratios) are examined within a "model". A theoretical model is constructed by the sociologist, who specifies that various effects be present or absent. The odds ratios within the theoretical model are set, using a computer programme, to certain values which within the model would mean the absence or presence of these effects. The frequencies generated within the model are then compared to the actual data by means of a chi square test of fit. The appropriate probability value is then calculated, i.e., the model is "tested" against reality. The model containing various effects is accepted or rejected on the basis of how well the expected frequencies in the model match the actual frequencies in the data. So the usual usage of the p value undergoes a reversal. For p values of less than, say, 0.05 the theoretical model is rejected as it differs "significantly" from the data. For p values higher than 0.05 the theoretical model may be accepted as fitting the data. Values of p near the stipulated borderline do of course suggest a near-fit and in such instances the model might be accepted or rejected on the grounds of theoretical rather than statistical relevance (Goodman, 1978).

There are two approaches in evaluating which model is the best fit. One may begin with the "saturated" model in which all possible effects are present, and delete higher order (three or more variable) interactions until the fit of the model to the data becomes unacceptable. Alternatively one may start with the simplest model possible and add interaction terms until "an acceptable fit is obtained which cannot be significantly improved by adding terms" (Knoke and Burke, 1980:7). We have chosen the latter method because it generates parsimonious models, and allows simplicity of presentation.

Within the log-linear model itself we will be examining a series of the simplest possible interactions, i.e., three variable interactions, in which the association between two variables is significantly affected by the impact of a third variable. As Goodman points out, "For a four way table there are 113 possible unsaturated log-linear models" (Stumpf, 1983:9). The mass of extra information generated by including a fourth variable in a particular model destroys the elegance of the three-way model, and in our specific case, an examination of each of our predictor variables in terms of matric rating and pass/fail was deemed more appropriate than a search for a complex model.

For each three variable interaction the variables pass/fail (W) and matric rating (X) were entered into the model together with a single variable related to either pass/fail or matric rating or both.

In addition to not constructing very complex models we will concentrate on the hierarchical aspects of the model, i.e., we will not be building a total model of effects within categories of different variables across other variables, nor will we examine the relative strength of these effects. We will be concentrating almost solely on discriminating between interactions, to achieve the purpose set at the beginning of the chapter. Due to the CHAID analysis already performed we have a reasonable idea of what is actually "happening" in the data and our variable list has been refined so that all our variables are grouped in terms of the variables pass/fail and matric rating, which are of specific interest to us

We shall be using the log-linear form of model, which according to Knoke and Burke "makes no distinction between independent and dependent variables, but is used to examine relationships among categories of variables by analysing expected cell frequencies" (Knoke and Burke, 1980:5). The logit model on the other hand demands that one specify which are the independent and dependent variables: which is something that, guided by additional information from theory about which variable(s) might plausibly be considered "prior" to others, we want to investigate rather than stipulate.

In other words, while evaluating the models generated by the log-linear method, the goals of the study will constantly be kept in mind. These are the identification of pertinent variables, and preferably ones which can actually be used, which will aid the detection of students who because of their "disadvantaged" background may be excluded from the university, even though these students have a fair chance of passing first year. We are searching for categories of respondents whose matric rating has been lowered due to the influence of factors which when removed or absent would have precipitated an improvement in academic performance in matric.

Not all variables significantly related to matric rating were included in log-linear analysis. Since we are in effect looking at the possible intervening effect of matric rating on pass/fail, only variables significantly related to both, or variables which theoretically should be related to both and were not when separately cross-tabulated, were selected for analysis.

A list of variables inserted into the log-linear model, together with their grouped categorizations as provided by CHAID, is presented below.

Table 6.2 Variables Included in Log-Linear Analysis

VARIABLE	CHAID CATEGORIES
Sex	(Male,q)/(Female)
Race	(Black)/(White,q)
Matric Exam	(q), (Provincial/Jmb)/ (NSC, Indian Affairs)
Aggregate Stds 8,9 Pre Matric Maths/ English, Std 8 And 9 Maths/English	(A,B,C)/(D,E,q)
Repetition of a Standard	(One)/(q,None,One +)
Index of Feelings About Study Habits	(Same,Less)/(Harder)
Parents Divorced	(q,No)/(Yes)
Parents Deceased	(q,No)/(Yes)
Financial Hard.	(q,Yes)/(No)
Education Parents	(q,Degree)/(No degree)

For the first table both the observed values and all possible fitted models will be reproduced, by way of illustration. All other tables will show only the observed values, the accepted model and any models which the accepted model had to be evaluated against. The outcome of log-linear analysis may also be represented graphically. The results of the first table will be presented in graphic form, while other tables will only be depicted graphically where relevant. In the headings, "*" stands for "by".

Table 6.3 Observed Values of Financial Hard. (A) * Pass/Fail (W) * Matric Rating (X)

HARDSHIP	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
Yes	Low	31	28	47	59
	Medium(q)	20	28	58	48
	High	6	46	88	52
	TOTAL	57	102	64	159
No	Low	53	70	56	123
	Medium(q)	38	127	76	165
	High	14	200	93	214
	TOTAL	105	397	79	502

The p-values for all possible models as tested against the above table of data are as follows. For example, "W,XA" hypothesizes that the data reflects an association between X and A, and between W and X, but not between W and A - i.e., that X intervenes between W and A.

MODELS	D.F.	CHI SQUARE	PROBABILITY
W	10	315.56	0.0
X	9	479.50	0.0
A	10	308.73	0.0
W,X	8	229.35	0.0
X,A	8	292.52	0.0
A,W	9	128.58	0.0
W,X,A	7	112.37	0.0
WX	6	205.80	0.0
WA	8	114.80	0.0
XA	6	282.64	0.0
W,XA	5	102.49	0.0
X,WA	6	98.59	0.0
A,WX	5	18.82	0.00
WX,WA	4	5.05	0.28
WA,XA	4	88.71	0.0
XA,WX	3	8.94	0.03
WX,WA,XA	2	1.08	0.58

When one looks down the p-values, it is possible almost immediately to eliminate most of the models. Only two theoretical models fit the data to an acceptable degree: (WX,WA) and (WX,WA,XA). The first model denotes that both financial hardship and matric rating are associated with pass/fail at university, but that financial hardship is not evidently associated with matric rating. The second model includes the latter link. In order to evaluate which of two models to choose as our final one we need to know if adding the extra term (XA) to the model (WX,WA) improves the fit of the model significantly.

Obviously any term added to the model will improve the fit in some way, and a saturated model (all effects present), will fit the best. We are attempting to explain the observed data in a simple, but adequate, manner. A compromise must be reached between a perfect fit and an acceptable fit. All effects in the theoretical model which, when excluded, do not significantly affect the fit of a model, are deemed not essential for an explanation of the data. In the model above, adding the effect (XA) - matric performance and financial hardship - to the model changes the chi square likelihood ratio by 3.97 for 2 degrees of freedom for which p is approximately 0.2. In other words, adding the term (XA) to the model (WX,WA) does not significantly improve the fit of the model. The accepted model is thus (WX,WA) and it may be presented diagramatically as follows (with theory and/or commonsense supplying the arrowed "direction of causality"):

$$\begin{tikzcd} & \text{pass/fail (W)} \\ \text{financial hardship (A)} \arrow[ur] & \\ & \text{matric rating (} \arrow[uu] \end{tikzcd}$$

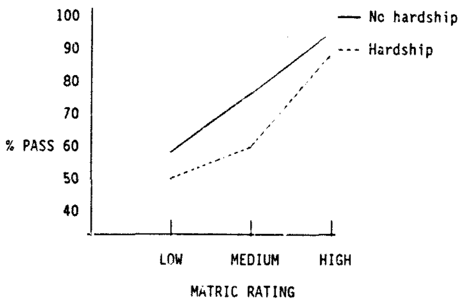
Graphically, the table may be shown as:

observed data in a simple, but adequate, manner. A compromise must be reached between a perfect fit and an acceptable fit. All effects in the theoretical model which, when excluded, do not significantly affect the fit of a model, are deemed not essential for an explanation of the data. In the model above, adding the effect (XA) - matric performance and financial hardship - to the model changes the chi square likelihood ratio by 3.97 for 2 degrees of freedom for which p is approximately 0.2. In other words, adding the term (XA) to the model (WX,WA) does not significantly improve the fit of the model. The accepted model is thus (WX,WA) and it may be presented diagramatically as follows (with theory and/or commonsense supplying the arrowed "direction of causality"):

$$\begin{tikzcd}
 & \text{pass/fail (W)} \\
\text{financial hardship (A)} \arrow[ur] & \\
 & \text{matric rating (X)} \arrow[uu]
\end{tikzcd}$$

Graphically, the table may be shown as:

Fig. 6.1 Percentage Pass vs. Matric Rating for Financial Hardship



Directed by the model that has been accepted, we look at the table of observed values and, attempting to think in "log-linear" terms, we find that the pass-to-fail ratio of respondents in the "low" category of matric rating who are enduring some form of financial hardship is about 1:1. In the medium category the ratio is about 2:3 (the actual ratio is 0.71) and in the high category about 3:5 (0.56). For respondents not experiencing financial hardship the ratio for the "low" category is 5:7 (0.75), the ratio of the medium category is about 4:13 (0.30) and the ratio in the high category is around 1:4 (0.26). The implication is clear. If one was forced to place a bet on pass/fail for various matric rating categories, the presence or absence of financial hardship would be good question to have an answer to, at every level of matric performance. (The fact that the graphs are the same shape reflects the lack of "pull" between financial hardship and matric rating.

Let us now turn to sex:

<u>Table 6.4 Observed Values of Sex (B) * Pass/Fail (W) * Matric Rating (X)</u>

SEX	MAT. RAT.	PASS Fail N	Pass N	% Pass	TOTAL
Female	Low	40	41	51	81
	Medium(q)	32	109	77	141
	High	17	178	91	195
	TOTAL	89	328	78	417
Male	Low	44	57	56	101
	Medium(q)	26	46	63	72
	High	3	68	95	71
	TOTAL	73	171	70	244

MODELS	D.F.	CHI SQUARE	PROBABILITY
XB,WX	3	6.52	0.08
WX,WB,XB	2	6.44	0.04

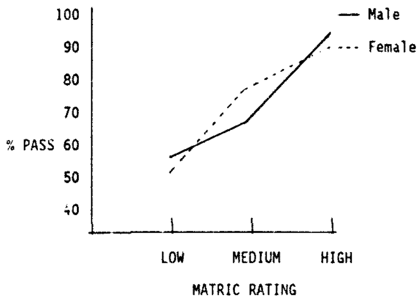
In the above table no model short of the saturated model is a good fit. Although the model containing the effects (XB,WX) does not significantly differ from the actual data the p value of 0.08 indicates a fit that is not really acceptable. So the saturated model, denoting an interaction, has to be chosen as the model for sex:

pass/fail (W)

sex (B)

matric rating (X)

Fig. 6.2 Percentage Pass vs. Matric Rating for Sex



Possible interpretations of an interaction are in principle symmetrical, although theory may indicate that some variables are best treated as antecedents and others as consequents. In this case, the relations between sex and pass may be said to differ across matric rating categories, or the relation between matric rating and university performance may be said to differ across sex. The latter is intuitively most plausible. It confirms information we already have from our cross-tabulations, but also tells us that the difference between the "curves" in the graph demands notice. For whatever reason, men with "middling" matric. pass first-year at a lower rate than women with comparative matrics, whereas at the extremes of matric performance the position is reversed.

What about race? Consider the following table.

Table 6.5 Observed Values for Race (C) * Pass/Fail (W) * Matric Rating (X)

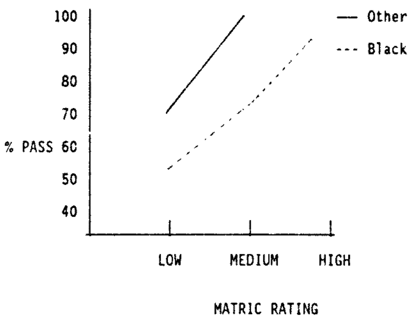
RACE	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
Black	Low	4	9	69	13
	Medium(q)	0	2	100	2
	High	0	0	-	0
	TOTAL	4	11	73	15
Other	Low	80	89	52	169
	Medium(q)	58	153	73	211
	High	20	246	92	266
	TOTAL	158	488	76	646

MODEL	D.F.	CHI SQUARE	PROBABILITY
XC,WX	3	2.64	0.45
WX,WC,XC	2	0.64	0.73

The log-linear model makes it clear that adding the term WC (pass/fail*race) to the model does not significantly improve the fit of the model to the data. The change in the chi square likelihood ratio with the extra term included is 2.00 for 1 degree of freedom - not significant. The final model (XC,WX) could be presented as:

$$\begin{tikzcd} & \text{pass/fail (W)} \\ \text{Black/other (C)} \arrow[r] & \text{matric rating (X)} \arrow[u] \end{tikzcd}$$

Fig. 6.3 Percentage Pass vs. Matric Rating for Race



Neither the effect between race and pass/fail, nor the interaction among the three variables, is strong enough to form a necessary part of the model. Blacks are passing at the same rate as the category "other" as the graph demonstrates, even though they are clustered at the lower end of the matric rating scale (13:2 vs about 9:25). However, the effect of matric rating on pass/fail is necessary for the explanation of the data.

We now present and interpret a series of log-linear analyses of the relations between other background variables on the one hand, and matric rating and pass/fail on the other.

Table 6.6 Observed Kinds of Matric Exam (D) * Pass/Fail (W) * Matric Rating (X)

MATRIC EX.	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
Provincial	Low	70	72	51	142
	Medium(q)	47	127	80	174
	High	17	225	93	242
TOTAL		134	424	80	558
NSC/Ind.Af	Low	8	15	65	23
	Medium(q)	4	10	71	14
	High	1	3	75	4
TOTAL		13	28	68	41
(q)	Low	6	11	65	17
	Medium(q)	7	18	72	25
	High	2	18	90	20
TOTAL		15	47	76	62

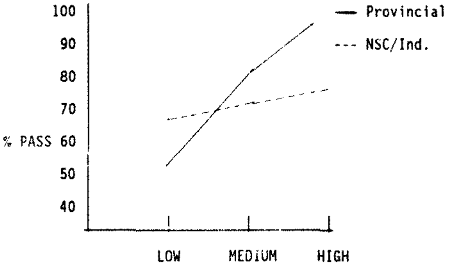
MODEL	D.F.	CHI SQUARE	PROBABILITY
XD,WX	6	4.03	0.67

pass/fail (W)

matric exam (D)

matric rating (X)

Fig. 6.4 Percentage Pass vs. Matric Rating for Matric Exam



The above model is the only one that fits the data. The observed table is difficult to interpret due to CHAID's separation of the "missing" category from the other categories of the variable matric exam. But it is fascinating to note that people with a NSC or Indian Affairs matric exam who fall into the low category of matric rating are passing at a much higher rate than provincial candidates in the same matric rating category, whereas the position is reversed for the medium and the high parts of the matric scale.

Table 6.7 Observed Values of Aggregate Std. 9 (E) * Pass/Fail (W) * Matric Rating (X)

AGG. 9	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
High	Low	31	37	54	68
	Medium(q)	35	87	71	122
	High	14	209	94	223
	TOTAL	80	333	81	413
Low	Low	53	61	54	114
	Medium(q)	23	68	75	91
	High	6	37	86	43
	TOTAL	82	166	67	248

MODEL	D.F.	CHI SQUARE	PROBABILITY
XE,WX	3	2.93	0.40
WX,WE,XE	2	2.85	0.24

Adding the term (WE) to the model (XE,WX) does not significantly improve the fit of the model (change in chi square likelihood ratio is 0.08 for 1 degree of freedom - not significant).

pass/fail (W)
↑
std. nine aggregate (E) → matric rating (X)

This kind of model, with school subject variable linking to matric rating, and matric rating linking to university pass, occurs in several more instances, as shown below: maths pre-matric, maths in std. 9, English pre-matric, and English in std. 9.

Table 6.8 Observed Values of Maths Pre-Matric (F) * Pass/Fail (W) * Matric Rating (X)

MATHS	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
High	Low	12	17	59	29
	Medium(q)	12	37	76	49
	High	9	125	93	134
	TOTAL	33	179	84	212
Low	Low	72	81	53	153
	Medium(q)	46	118	72	164
	High	11	121	92	132
	TOTAL	129	320	71	449

MODEL	D.F.	CHI SQUARE	PROBABILITY
XF,WX	3	0.91	0.85

pass/fail (W)

pre-matric maths (F)

matric rating (X)

Table 6.9 Observed Values of Std. Nine Maths (G) * Pass/Fail (W) * Matric Rating (X)

AGG. 9	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
High	Low	19	17	47	36
	Medium(q)	13	50	79	63
	High	10	158	94	168
	TOTAL	42	225	84	267
Low	Low	65	81	55	146
	Medium(q)	45	105	70	150
	High	10	88	90	98
	TOTAL	120	274	70	394

MODEL	D.F.	CHI SQUARE	PROBABILITY
XG,WX	3	4.38	0.22

pass/fail (W)
↑
maths std. nine (G) → matric rating (X)

Table 6.10 Observed Values of Pre-Matric English (H) * Pass/Fail (W) * Matric Rating (X)

ENGLISH	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
High	Low	34	49	59	83
	Medium(q)	34	94	73	128
	High	14	213	94	227
	TOTAL	82	356	81	438
Low	Low	50	49	49	99
	Medium(q)	24	61	72	85
	High	6	33	85	39
	TOTAL	80	143	64	223

MODEL	D.F.	CHI SQUARE	PROBABILITY
XH,WX	3	5.09	0.17
WX,WH,XH	2	2.25	0.32

The change in the chi square likelihood ratio is 2.84 for one degree of freedom, i.e., p = 0.1. So adding the term (WH) does not significantly improve the model.

pass/fail (W)

pre-matric English (H) → matric rating (X)

<u>Table 6.10 Observed Values of Pre-Matric English (H) * Pass/Fail (W) * Matric Rating (X)</u>

ENGLISH	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
High	Low	34	49	59	83
	Medium(q)	34	94	73	128
	High	14	213	94	227
	TOTAL	82	356	81	438
Low	Low	50	49	49	99
	Medium(q)	24	61	72	85
	High	6	33	85	39
	TOTAL	80	143	64	223

MODEL	D.F.	CHI SQUARE	PROBABILITY
XH,WX	3	5.09	0.17
WX,WH,XH	2	2.25	0.32

The change in the chi square likelihood ratio is 2.84 for one degree of freedom, i.e., $p = 0.1$. So adding the term (WH) does not significantly improve the model.

pass/fail (W)
↑
pre-matric English (H) → matric rating (X)

Table 6.11 Observed Values of English Std. 9 (I) * Pass/Fail (W) * Matric Rating (X)

ENG. 9	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
High	Low	46	56	55	102
	Medium(q)	40	102	72	142
	High	15	216	94	231
	TOTAL	101	374	79	475
Low	Low	38	42	53	80
	Medium(q)	18	53	75	71
	High	5	30	86	35
	TOTAL	61	125	67	186

MODEL	D.F.	CHI SQUARE	PROBABILITY
XI,WX	3	2.52	0.47

$$\text{std. nine English (I)} \longrightarrow \text{matric rating (X)} \longrightarrow \text{pass/fail (W)}$$

The four preceding tables show that the measured variables associated with school record affected matric rating. This is a to be expected, since both are measures of the same thing. What is relevant is that none of the school variables on their own had enough influence on pass/fail at university for the effect to be included in any of the models above. Indeed the only one which came close to affecting pass/fail in a model was matric English.

We now turn to other factors from schooling: failing a year, and study habits.

Table 6.12 Observed Values of Repetition Standard (J) * Pass/Fail (W) * Matric Rating (X)

REPEAT	MAT. RAT.	PASS Fail N	Pass N	% Pass	TOTAL
One	Low	46	56	55	102
	Medium(q)	40	102	72	142
	High	15	216	94	231
	TOTAL	101	374	79	475
None	Low	38	42	53	80
One +	Medium(q)	18	53	75	71
	High	5	30	86	35
	TOTAL	61	125	68	186

MODEL	D.F.	CHI SQUARE	PROBABILITY
WJ,WX	3	2.52	0.47

pass/fail (W)

repetition of a std. (J) → pass/fail (W) ← matric rating (X)

matric rating (X)

Students who repeated one standard at school have lower pass rates at university than students who did not repeat a standard or who repeated more than one standard. At school, however, this variable did not influence the matric performance of respondents significantly. The difference that is uncovered between this model and the previous four;

concerning pre-matric and std. 9 scores, is a neat vindication of the use of the log-linear approach.

Table 6.13 Observed Values of Index of Study (K) * Pass/Fail (W) * Matric Rating (X)

STUDY	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
High	Low	51	56	52	107
	Medium(q)	34	117	77	151
	High	16	191	92	207
	TOTAL	101	364	78	'65
Low	Low	33	42	56	75
	Medium(q)	24	38	61	62
	High	4	55	93	59
	TOTAL	61	135	69	196

MODEL	D.F.	CHI SQUARE	PROBABILITY
XK,WX	3	5.90	0.12

pass/fail (W)

index of study (K) → matric rating (X) → pass/fail (W)

matric rating (X)

It turns out that the effect of respondents' reported study habits are not directly necessary for an understanding of university pass/fail rates. The effect is, however, necessary for an understanding of differences in matric rating scores.

Let us now turn from possible indicators within school performance to factors in the respondents' home situation which may bear on subsequent university performance: whether parents are divorced or deceased, and their education levels.

Table 6.14 Observed Values of Parents Divorced (L) * Pass/Fail (W) * Matric Rating (X)

DIVORCE	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
Yes	Low	17	20	54	37
	Medium(q)	5	17	77	22
	High	2	27	93	29
	TOTAL	24	64	73	88
No	Low	67	78	54	145
	Medium(q)	53	138	72	191
	High	18	219	92	237
	TOTAL	138	435	76	573

MODEL	D.F.	CHI SQUARE	PROBABILITY
L,WX	5	10.3	0.07
XL,WX	3	0.28	0.96
WX,WL,XL	2	0 16	0.92

The change in the likelihood ratio is 10.02 for 2 degrees of freedom, p = 0.01. In other words, although the model containing only the effect WX fits the data, adding the effect XL significantly improves the fit of the model. However, adding the term (WL) to the model does not significantly improve the fit of the model. So the final model (XL,WX) may be presented as:

pass/fail (W)

parents divorced/separated (L) → matric rating (X) → pass/fail (W)

In sum a respondent's matric rating, but not directly university pass/fail, is affected in an adverse way by the divorce or separation of the respondent's parents.

Table 6.15 Observed Values of Parents Deceased (M) * Pass/Fail (W) * Matric Rating (X)

DECEASED	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
No	Low	77	80	51	157
	Medium(q)	52	150	74	202
	High	18	228	93	246
	TOTAL	147	458	76	605
Yes	Low	7	18	72	25
	Medium(q)	6	5	45	11
	High	2	18	90	20
	TOTAL	15	41	73	56

MODEL	D.F.	CHI SQUARE	PROBABILITY
XM,WX	3	8.02	0.04
WX,WM,XM	2	7.90	0.01

No model effectively fits the data so a saturated model must be assumed, i.e., an interaction is present. It seems that among students of middle matric rating the pass rate is much lower for students whose parents are deceased than for those whose parents are alive. Among students of low matric rating, by contrast the pass rate is higher among those whose parents are deceased. But we also note that the actual proportion of students in the middle category of the matric rating scale is much lower for students whose parents are deceased (N = 11). So we conclude that the low sample size of the "parents deceased" variable does not justify a display of the model.

Table 6.16 Observed Values of Father's Level of Education (O) * Pass/Fail (W) * Matric Rating (X)

EDUCATION	MAT. RAT.	PASS			TOTAL
		Fail N	Pass N	% Pass	
No Degree	Low	59	52	47	111
	Medium(q)	44	91	67	135
	High	15	136	90	151
TOTAL		118	279	70	397
Degree	Low	25	46	64	71
	Medium(q)	14	64	82	78
	High	5	110	96	115
TOTAL		44	220	83	264

MODEL	D.F.	CHI SQUARE	PROBABILITY
WX,WO	4	1.46	0.83
WX,WO,XO	2	0.06	0.97

The change in the chi square likelihood ratio is 1.4 for 2 degrees of freedom, for which p is approximately 0.5. So, adding adding the term XO does not significantly improve the