Clustering students by their subject choices in the Learning Curves project

Hilary Ferral

NEW ZEALAND COUNCIL FOR EDUCATIONAL RESEARCH

TE RÜNANGA O AOTEAROA MÖ TE RANGAHAU I T E MÄTAURANGA

WELLINGTON

2005

New Zealand Council for Educational Research
P O Box 3237
Wellington
New Zealand

Table of Contents

Tables

Figures

1. Introduction

The purpose of this paper is to describe and report findings from the clustering process applied to the Learning Curves 2004 data. It is of a technical nature, and is designed to complement the third report from the Learning Curves project (Hipkins & Vaughan, with Beals, Ferral, & Gardiner, 2005).

The analysis is exploratory and seeks to isolate patterns in the data related to students' subject choices. We wished to discover whether the data showed relationships between subject choices and the student demographics we collected – namely students' school, gender, and self-defined ethnic group. Students were grouped (clustered) according to their reported subject choices, i.e. those with similar ranges of subject choices were grouped together. We then cross-tabulated these patterns of subject choice with our three demographic variables.

In the Learning Curves project the scope for this enquiry is somewhat constrained. First, as this analysis was not part of the original brief, neither the questionnaire nor the sample design are quite ideal. In particular, the demographic information we have about the students is limited to the three variables already mentioned. Second, the Learning Curves sample is not a random sample, so we are unable to infer anything beyond the six case study schools. Third, the level of non-response gives cause for concern. The amount of resultant bias in the sample is unknown but likely to be non-trivial. These constraints notwithstanding, the analysis did show some interesting results and points to possible further research in this area.

The paper is organised as follows: Section 2 gives enough technical background to understand the processes used. Section 3 is a "what we did" section giving a description of the data, and preparation for the cluster analysis. It also includes a short discussion on the limitations of the data with reference to this analysis. Section 4 sets out the results with comments and observations. In the final section (5) a summary of the findings, conclusions, and pointers to further research are presented.

2. Clustering – technical background

Successful clustering requires three fundamental decisions to be made. We must first establish distance or similarity measure to distinguish how “close” observations are to one another. Second, a suitable clustering algorithm must be chosen – there is a very wide range of choices – to group the data, and third we need to choose a sensible way to measure the “distance” between intermediate clusters during the clustering process.

2.1 Distance and similarity measures

Clustering begins by establishing a measure of “similarity” between observations (students in this case) with respect to their subject choice. Two students who take exactly the same subjects as each other are completely similar. Other students, whose subject choices are not all the same, are less similar. The Learning Curves subject choice data is represented by binary variables. Each subject forms one variable, which equals 1 if a student is taking a subject, and 0 otherwise.

There are a number of possible similarity measures to choose from. In this instance it is appropriate to use the Jaccard similarity coefficient (Sneath, 1957), which is constructed as follows.

Suppose two students have the following subject choice profile:

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7
Student 1	1	1	0	0	1	0	1
Student 2	0	1	1	0	1	1	0

We may then create a cross-tabulation of where the students’ subject choice agrees or disagrees.

		Student2	
		1	0
Student 1	1	A = 2	B = 2
	0	C = 2	D = 1

The Jaccard similarity coefficient is then calculated as:

$$J = \frac{A}{A+B+C} = \frac{2}{6} = \frac{1}{3}$$

In other words, the coefficient can be described as the proportion of positive matches with respect to the sum of possibilities. Note that the Jaccard coefficient ignores those instances where neither student is taking a subject. Other similarity measures take the negative matches into account, but in our case our interest is in which subjects students ***are*** taking, rather than the subjects they are ***not*** taking, so this is an appropriate measure to use.

Suppose the students in the example above had had exactly the same choice of subjects, i.e. A = 7, B = 0, and C = 0, then we have $J = 1$. On the other hand if the two students do not match on any subjects then we get $J = 0$. So the range of J is from 0 to 1, with coefficients close to 0 indicating little similarity and coefficients close to 1 indicating a high degree of similarity between observations.

Dissimilarity coefficients (or distance measures) can be calculated as $1 - J$. Most software packages will accept similarities or dissimilarities with equal ease, so it is up to the user to decide which is most appropriate. In the Learning Curves case it makes sense to talk about clustering similar students, so we consider similarity coefficients.

A similarity matrix containing similarity coefficients for ***all*** distinct pairs of observations must be calculated – that is, $\frac{n(n-1)}{2}$ coefficients, where n is the number of observations. The matrix is then handed to the clustering procedure (see next section). We used the SAS (SAS Institute Inc, 1999–2001) macro **`%distance`** to calculate the similarity matrix for our data.

2.2 Clustering algorithms

Many varied algorithms for clustering observations are available. Everitt, Landau, and Leese (2001) and Kaufmann and Rousseeuw (1990) both give full accounts of clustering techniques. The clustering algorithms explored for the Learning Curves data are of the hierarchical type. Hierarchical algorithms can be split into two general methods. The divisive method begins with the data in one large cluster and makes stepwise divisions in the data to form clusters, ending up with n clusters of individual observations. The other approach is the agglomerative method, which begins with n clusters (of individual observations) and joins the most similar observations (or small clusters) in a stepwise procedure, this time ending up with one large cluster containing all the data. Details of the dividing or joining steps are recorded by the algorithm. We used the SAS (SAS Institute Inc, 1999–2001) procedure **`proc cluster`** to cluster observations with an agglomerative algorithm.

2.3 Dendrograms

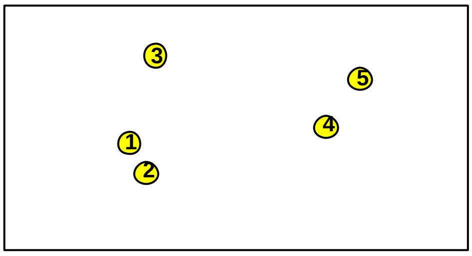
The SAS procedure **`proc cluster`** produces dendrograms to help analyse the clusters. Dendrograms are visual representations of the clustering process. For example, suppose we have a similarity matrix, *S*.

Figure 1 Example similarity matrix

$$
S = \begin{array}{c|ccccc}
 & 1 & 2 & 3 & 4 & 5 \\
\hline
1 & - & - & - & - & - \\
2 & 0.90 & - & - & - & - \\
3 & 0.70 & 0.65 & - & - & - \\
4 & 0.40 & 0.40 & 0.55 & - & - \\
5 & 0.30 & 0.30 & 0.50 & 0.80 & -
\end{array}
$$

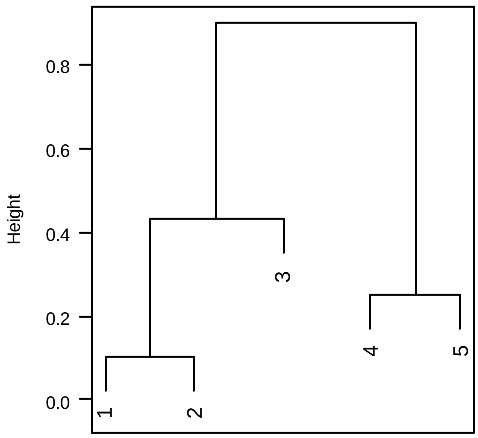
Visually this matrix could approximately describe the following situation

Figure 2 Visual representation of similarity matrix (Figure 1)



which produces the following very simple dendrogram.

Figure 3 Dendrogram relating to similarity matrix (Figure 1)



From Figure 2 we see that observations 1 and 2 are the most similar, so they are the first to be joined. Observations 4 and 5 are also very similar so they are joined next. In the third step the small cluster of 1 and 2 is joined to observation 3, and finally the cluster with observations 1, 2, and 3 is joined with the cluster containing observations 4 and 5, making one grand cluster containing all observations.

2.4 Creating clusters

A question which arises naturally from the previous paragraph is how to measure the distance between intermediate clusters formed by successive steps of the algorithm. Should we measure between the closest members of each cluster? The members furthest away from each other? From centre to centre? And if centre to centre, how should we define the centre of an irregularly shaped cluster? And so on… The most successful methods measuring inter-cluster distance for the Learning Curves data were the "flexible" method developed by Lance and Williams (1967), and the method attributed to Ward (1963). These methods were "successful" in that they produced more clearly defined clusters of more even size than some other methods.

Another question to be answered is "At what point should the algorithm be stopped?" That is, how many clusters should we create? The dendrograms are useful here. If we were to draw a horizontal line across the dendrogram at some arbitrary height, say 0.6 in our example, the line will cross two vertical lines of the dendrogram, giving us two clusters, one containing observations 1, 2, and 3, the other containing observations 4 and 5. As to which is the "right" or the "best" height at which to cut the dendrogram, there are, unfortunately, no definitive answers. We need to base these decisions on current investigations and, if available, supporting evidence from other studies.

Faced with exploring a set of data with a view to finding interesting groupings, there are choices to be made in the process. These choices all influence the results to one degree or another. Whether they are the "right" choices, or whether one choice is "better" than another are not easy judgements to make.

Everitt et al. (2001) comment "It is generally impossible a priori to anticipate what combinations of variables, similarity measures and clustering techniques are likely to lead to interesting and informative classifications." A pragmatic approach is recommended. Everitt's advice is to explore and compare many different (appropriate) methods. Similarity in the results from different methods gives more confidence that the patterns genuinely exist in the data. Results that appear to be sensitive to the method used inspire less confidence.

3. Processing the Learning Curves data

3.1 Data description

For a full description of the Learning Curves data see Hipkins et al. (2005). The clustering analysis required data about the subjects students had chosen to take, and some demographic variables.

Students were asked to tick off subjects they were doing from a prepared list of all Year 11, 12, and 13 subjects available at their school. Although the subject lists were school-specific there was commonality amongst subjects across schools. For example, Mathematics 101, Mathematics, Mathematics MAT, Full NCEA Mathematics, and Mathematics Level 1 are all names for the traditional Year 11 mathematics course. We gave subjects common names across all schools, so we could analyse all schools together. We treated each year level separately to cater for the differences in subject choice practice between Year 11, 12, and 13. For example, at Year 11 some form of English, mathematics, and science is compulsory in most secondary schools leaving comparatively limited opportunities for genuine choice. At Year 12 more choices exist for most students, although English is commonly a compulsory subject. Year 13 students have the most opportunities for genuine subject choice. Only one of our Learning Curves schools made English a compulsory subject at Year 13.

Each subject is recorded as a binary variable:

$$X_{ij} = \begin{cases} 1 & \text{if student } i \text{ takes subject } j \\ 0 & \text{otherwise} \end{cases}$$

If a student is taking a subject it is recorded as a 1, otherwise a 0 is recorded.

Students were also asked to indicate their gender, and self-defined ethnic group(s). For comparing clustering results to ethnicity we used the SNZ[1] prioritising scheme for ethnic groups. We have used the groupings:

- Mäori;
- Pacific;
- Asian;

[1] Statistics New Zealand ethnic classification level 1. See http://www.stats.govt.nz/census/2001-born-overseas/explanatory-notes.htm for further information

- Päkehä; and
- Other/unknown/missing.

Students who identified multiple groups were assigned to one of the groups above. The groups are listed in descending order of priority.

We also have a school identification number for each student.

3.2 Data preparation

Tractable clusters depend on certain characteristics in the data. Ideally we should have many more observations than variables. Once the students who had offered no information about their subject choices had been removed, we restricted the Year 11 clustering to Year 11 subjects and Year 12 traditional mathematics since, apart from the mathematics, there were very few students taking subjects at another year level. Years 12 and 13 were similarly restricted to subjects within their own year level.

We then took the pragmatic step of combining some subjects under one umbrella. For example, we grouped the subjects Technology (Soft Materials), Food Technology, Technology (Hard/Soft Materials), Technology (Hard Materials) together under "Practical Technology". Where there is sufficient similarity between subjects to do this "collapsing", it means that we can make use of the data rather than having to discard it because it is too fragmented. Subjects taken by less than 3 percent of students were eliminated. Information about these subjects simply adds noise to an already noisy environment, so are better left out. Overall, after collapsing and eliminating, the Year 11 subject list was reduced from 48 subjects to 38, the Year 12 list from 57 to 43 subjects, and the Year 13 list from 54 to 41 subjects.

3.3 Non-response issues

In some schools the response rate was poor. This is likely to cause some bias in the results. It depends upon the pattern of non-response what this bias will be. It is possible that only specific subject classes answered the questionnaire at some schools or that specific subject classes are missing. In this situation we will only pick up part-information for a whole school. We need to be mindful of this when looking at the results.

3.4 Process

We used the Jaccard similarity coefficient (see previous section) to measure similarity between students' subject choices (SAS macro **`%distance`**). An agglomerative clustering algorithm was used to cluster the observations (SAS **`proc cluster`**). The most successful clustering methods were the “flexible” method developed by Lance and Williams (1967), and the method attributed to Ward (1963). Both algorithms produced identical clusters in terms of subject choice, with almost identical observations within clusters. The results reported are those from the flexible method.

4. Results

4.1 Year 11 results

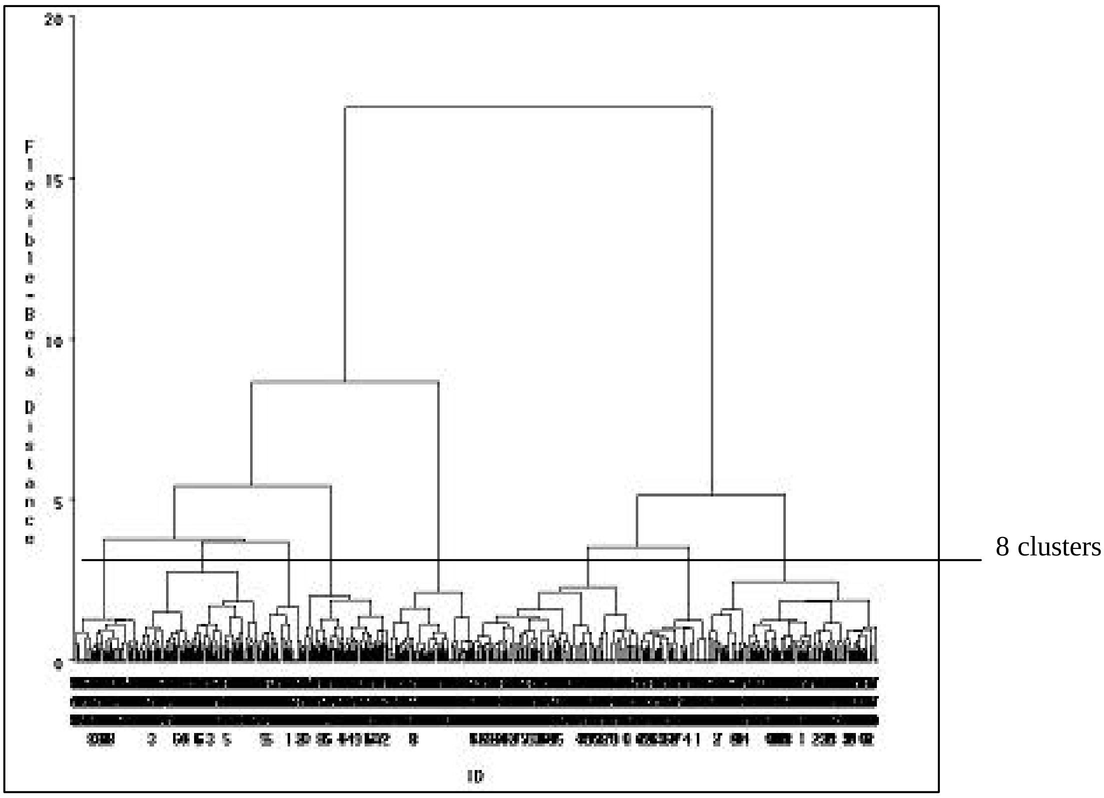
Subjects were chosen to characterise clusters when membership was more than 20 percent above the expected membership. For example, 80.5 percent of the Year 11 students are taking traditional English overall (see Table 1). Clusters 1, 2, 3, 4, and 6 are distinct in that **all** students in these clusters are taking traditional English. Therefore these clusters are characterised by the subject traditional English. A further example: 60.5 percent of all Year 11 students in our dataset are taking traditional mathematics, but clusters 1, 3, 4, and 8 have very nearly all students taking traditional mathematics. This means that these clusters are characterised by traditional mathematics. Observe also that clusters 2, 5, and 7 have no students taking traditional mathematics, but that they are represented very strongly by students taking alternative mathematics. Cluster 6 is characterised by students taking traditional mathematics at a Year 12 level.

Table 1 **Year 11 clusters**

	Overall	CLUS1	CLUS2	CLUS3	CLUS4	CLUS5	CLUS6	CLUS7	CLUS8
Traditional English	80.49%	100.0%	100.0%	100.0%	100.0%	0.00%	100.0%	74.49%	52.86%
Contextually-focused English	10.82%	0.00%	0.00%	0.00%	0.00%	61.43%	0.00%	.22%	24.29%
Media Studies	3.35%	7.41%	1.89%	4.32%	1.79%	0.00%	2.86%	2.04%	1.43%
ESOL	4.12%	1.48%	0.00%	0.00%	0.00%	14.29%	0.00%	3.06%	17.14%
Traditional Mathematics	60.52%	99.26%	0.00%	100.0%	100.0%	0.00%	0.00%	0.00%	97.14%
Alternative Mathematics	33.99%	0.00%	100.0%	0.72%	1.79%	100.0%	0.00%	97.96%	2.86%
Accounting	7.62%	.11%	3.77%	.51%	3.57%	1.43%	2.86%	4.08%	12.86%
Traditional Science	76.98%	100.0%	100.0%	100.0%	100.0%	0.00%	100.0%	69.39%	27.14%
Alternative Science	12.96%	0.00%	0.00%	0.00%	0.00%	77.14%	0.00%	19.39%	17.14%
Agriculture/Horticulture	8.69%	1.48%	7.55%	7.19%	0.00%	14.29%	0.00%	14.29%	24.29%
Health & Physical Education	54.42%	99.26%	100.0%	0.00%	100.0%	51.43%	40.00%	29.59%	50.00%
Health & Lifeskills	14.94%	22.22%	7.55%	6.47%	5.36%	25.71%	0.00%	13.27%	30.00%
Recreation	4.12%	0.74%	3.77%	13.67%	0.00%	0.00%	2.86%	2.04%	2.86%
Food & Nutrition	8.99%	.11%	20.75%	6.47%	7.14%	2.86%	0.00%	13.27%	7.14%
Geography	.13%	8.89%	3.77%	14.39%	23.21%	1.43%	8.57%	17.35%	7.14%
History	22.10%	8.15%	.32%	25.18%	100.0%	4.29%	45.71%	12.24%	8.57%
Economics	8.84%	10.37%	1.89%	10.07%	19.64%	0.00%	17.14%	6.12%	8.57%
Economics & Accounting	6.55%	3.70%	0.00%	8.63%	1.79%	2.86%	34.29%	10.20%	1.43%
European Languages	6.40%	5.93%	0.00%	5.76%	26.79%	1.43%	14.29%	1.02%	5.71%
Te Reo Mäori	3.81%	3.70%	9.43%	1.44%	8.93%	2.86%	0.00%	3.06%	4.29%
Practical Technology	22.10%	25.93%	16.98%	26.62%	0.00%	27.14%	14.29%	23.47%	24.29%
Graphics and Design	18.45%	23.70%	9.43%	30.22%	1.79%	5.71%	25.71%	16.33%	17.14%
Information Management	18.14%	17.04%	.32%	22.30%	5.36%	17.14%	5.71%	21.43%	30.00%
Computer Studies	14.63%	7.41%	18.87%	15.11%	1.79%	30.00%	31.43%	16.33%	8.57%
Visual Arts	17.38%	19.26%	1.89%	25.90%	0.00%	14.29%	5.71%	21.43%	25.71%
Music	10.98%	5.93%	13.21%	17.27%	14.29%	7.14%	22.86%	3.06%	12.86%
Drama	12.20%	17.04%	7.55%	17.99%	10.71%	10.00%	0.00%	10.20%	7.14%
Transition	13.41%	26.67%	15.09%	0.72%	16.07%	24.29%	0.00%	10.20%	10.00%
Technology – Vocational Pathways	12.96%	2.22%	20.75%	9.35%	1.79%	27.14%	.43%	25.51%	12.86%
Traditional Mathematics (Yr12 level)	5.34%	0.00%	0.00%	0.00%	0.00%	0.00%	100.0%	0.00%	0.00%

The following dendrogram (Figure 4) shows the hierarchical structure of the clusters. An optimal number of clusters can be chosen by “cutting” the tree at a certain height. In general, cutting a tree where the difference in height between successive steps of the procedure is comparatively large is a good idea, ensuring a clear distinction between clusters. Additionally, relatively even sized clusters will render more robust comparisons between clusters and other variables. With these points in mind we decided to use eight clusters for the Year 11 students.

Figure 4 Dendrogram for Year 11 clusters



A table of subject group cluster characteristics follows. To make sense of this table we can broadly say that “most students in a particular cluster are taking most subjects which characterise that cluster”. The columns at the bottom of the table contain initial overall observations about the nature of the subjects taken by the students in each cluster.

Table 2 **Year 11 cluster characteristics**

Cluster1 *n = 135*	Cluster2 *n = 53*	Cluster3 *n = 139*	Cluster4 *n = 56*	Cluster5 *n = 70*	Cluster6 *n = 35*	Cluster7 *n = 98*	Cluster8 *n = 70*
Traditional English Traditional Mathematics Traditional Science Media Studies Accounting Health & Physical Education Health & Lifeskills Food & Nutrition Graphics & Design Drama Transition	Traditional English Alternative Mathematics Traditional Science Health & Physical Education Food & Nutrition Te Reo Mäori Computer Studies Music Technology – Vocational Pathways	Traditional English Traditional Mathematics Traditional Science Media Studies Accounting Recreation Geography Economics & Accounting Practical Technology Graphics & Design Information Management Visual Arts Music Drama	Traditional English Traditional Mathematics Traditional Science Health & Physical Education Geography History Economics European Languages Te Reo Mäori Music	Contextually-focused English Alternative Mathematics Alternative Science ESOL Agriculture/ Horticulture Health & Lifeskills Practical Technology Computer Studies Transition Technology – Vocational Pathways	Traditional English Traditional Mathematics at Year 12 level Traditional Science History Economics Economics & Accounting European Languages Graphics & Design Computer Studies Music	Alternative Mathematics Alternative Science Agriculture/ Horticulture Food & Nutrition Geography Economics & Accounting Visual Arts Technology – Vocational Pathways	Contextually-focused English Traditional Mathematics Alternative Science ESOL Accounting Agriculture/ Horticulture Health & Lifeskills Information Management Visual Arts
Overall descriptions of subjects which characterise the clusters							
Traditional core subjects Other practical	Traditional English & science Alternative mathematics Other practical	Traditional core subjects Other mixed practical/ academic	Traditional core subjects Other academic	Alternative core subjects ESOL Other practical	Traditional core subjects with accelerated mathematics Other academic	Alternative mathematics and science Other mixed academic and practical	Traditional mathematics Alternative English and science ESOL Other practical

Having established clusters of students based on their subject choices alone, we were interested to discover whether the clusters are associated with school, ethnic group, or gender. In other words are the clusters into which students naturally fall (based on subject choice alone) school-specific, ethnicity-specific, or gender-specific, or combinations of these?

Table 3 shows how the Year 11 students fall across the clusters with respect to their school. The percentages in the body of the table show the proportion of students in each school for a particular cluster. For example, 37.78 percent of the students in Cluster 1 are attending City School A and 20 percent of Cluster 1 students are attending City School B and so on. The column on the far right shows the proportion of students overall who attend the separate schools. That is, 20.27 percent of students in the Year 11 cohort attend City School A, and 15.4 percent of the Year 11 cohort attend Town School F. Comparing these two percentages gives us an idea about where clusters are over- or under-represented by the various schools. We see that City School A is over-represented in Cluster 1 and Cluster 4, and also that Cluster 1 is over-represented by students from City Schools A and B, and Town School F. The percentages in the "total" row show the proportion of the whole cohort that belong to each of Clusters 1 to 8.

Note that sample sizes vary between tables because of missing data for the responses in question.

A chi-square test of association between school and cluster produces a p-value of <0.0001. This indicates that the data support a hypothesis of association between school and cluster. Each school tends to dominate in two or three of the eight clusters, and each cluster is dominated by two or three schools. City School A dominates in Cluster 4, which is characterised by more academic subjects. City School C predominantly populates Clusters 3, 5, and 8, with the most marked membership in Cluster 5 which is characterised by alternative core subjects, practical subjects, and ESOL. Cluster 6 (characterised by more academic subjects and accelerated mathematics courses) has dominant membership from Town School E.

Table 3 **School by Year 11 cluster**

School ↓		CLUSTER								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Overall**
A	*n*	**51**	10	1	**38**	10	1	8	14	133
	%	**37.78**	18.87	0.72	**67.86**	14.29	2.86	8.16	20.00	20.27
B	*n*	**27**	6	**38**	10	2	5	7	**15**	110
	%	**20.00**	11.32	**27.34**	17.86	2.86	14.29	7.14	**21.43**	16.77
C	*n*	2	0	**25**	1	**16**	4	12	**13**	73
	%	1.48	0.00	**17.99**	1.79	**22.86**	11.43	12.24	**18.57**	11.13
D	*n*	18	**14**	17	1	8	4	**20**	8	90
	%	13.33	**26.42**	12.23	1.79	11.43	11.43	**20.41**	11.43	13.72
E	*n*	9	**19**	23	3	**26**	**21**	**41**	7	149
	%	6.67	**35.85**	16.55	5.36	**37.14**	**60.00**	**41.84**	10.00	22.71
F	*n*	**28**	4	**35**	3	8	0	10	13	101
	%	**20.74**	7.55	**25.18**	5.36	11.43	0.00	10.20	18.57	15.40
Total	*n*	135	53	139	56	70	35	98	70	656
	%	20.58	8.08	21.19	8.54	10.67	5.34	14.94	10.67	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	307.13
Df	35
p-value	<.0001
Sample size	656

Table 4 shows a comparison of ethnic group with cluster. There are more Asian students than expected in Cluster 8, Mäori students in Clusters 2, 5, and 8, Pacific students in Clusters 5, 7, and 8, and Päkehä students in Clusters 3 and 4. These deviations from the expected frequencies amount to a significant chi-square statistic indicating an association between ethnic group and subject choice. The separation between the Päkehä group and other ethnic groups is distinctive here. The apparent association between ethnic group and cluster could be confounded by the school effect observed above. That is, if one or more schools has a particular predominance of any one ethnic group, we have no way of telling whether the association with subject choice groups is due to ethnic group or school. Log-linear models which might be able to isolate these effects are discussed later.

The "other/missing" ethnic group was removed for this table.

Table 4 **Ethnic group by Year 11 cluster**

Ethnic Group ↓		CLUSTER								
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Overall**
Asian	*n*	8	0	6	3	5	2	4	**10**	38
	%	6.90	0.00	4.62	6.00	8.62	5.88	4.4	**16.95**	6.45
Mäori	*n*	13	**11**	15	4	**14**	4	16	**13**	110
	%	11.21	**22.00**	11.54	8.00	**24.14**	11.75	17.58	**22.03**	15.31
Pacific	*n*	7	3	3	2	**9**	3	**12**	**9**	48
	%	6.03	6.00	2.31	4.00	**15.52**	8.82	**13.39**	**15.25**	8.15
Päkehä	*n*	88	36	**106**	**41**	30	25	59	27	412
	%	75.86	72.00	**81.54**	**82.00**	51.72	73.53	64.84	45.76	70.07
Total	*n*	116	50	130	50	58	34	91	59	588
	%	19.73	8.50	22.11	8.50	9.86	5.78	15.48	10.03	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	55.99
Df	21
p-value	<.0001
Sample size	588

(Small cell sizes may compromise the reliability of the test for this table.)

A gender analysis (Table 5) shows that male students dominate Clusters 6 and 7, while female students dominate in Clusters 1 and 4. This could also be a (partly) school-driven effect. We have already observed the one single-sex girls' school predominating in Clusters 1 and 4, and the one single-sex boys' school in the dataset predominating in Clusters 6 and 7.

Table 5 **Gender by Year 11 cluster**

Gender ↓		CLUSTER								
		1	2	3	4	5	6	7	8	Overall
Male	*n*	42	30	77	8	41	**28**	**65**	34	325
	%	32.56	56.60	57.89	14.55	61.19	**80.00**	**66.33**	51.51	51.10
Female	*n*	**87**	23	56	**47**	26	7	33	32	311
	%	**67.44**	43.40	42.11	**85.45**	38.81	20.00	33.67	48.48	48.90
Total	*n*	129	53	133	55	67	35	98	66	636
	%	20.28	8.33	20.91	8.65	10.53	5.50	15.41	10.38	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	73.79
Df	7
p-value	<.0001
Sample size	636

4.2 Year 12 Results

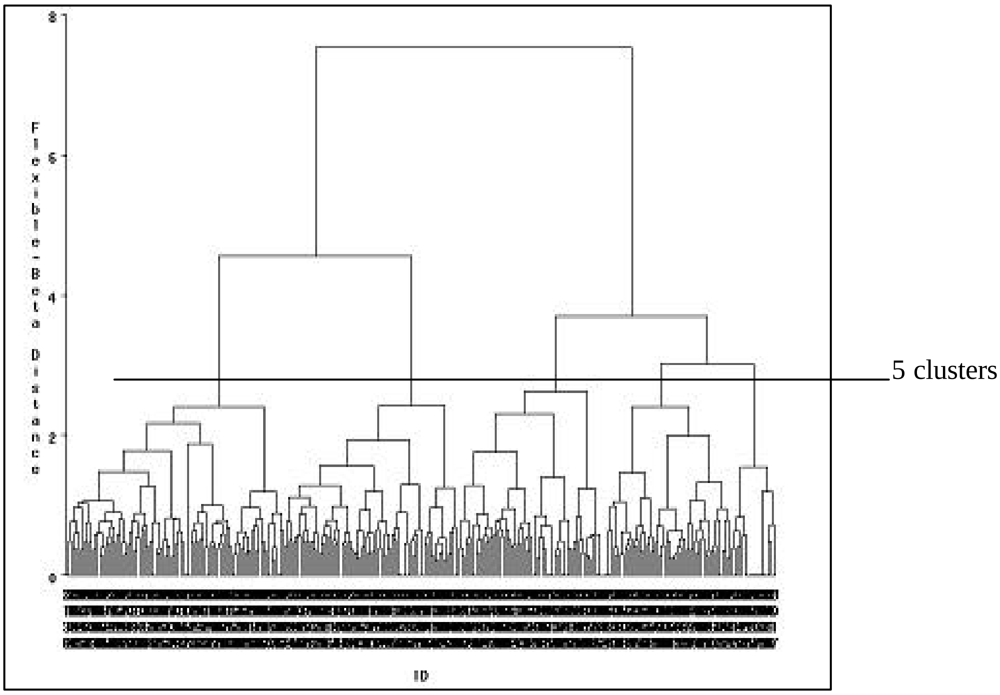
The Year 12 subject data was processed in a similar manner to the Year 11 data. Table 6 gives an overview of cluster characteristics at Year 12.

Table 6 **Year 12 clusters**

	Overall	CLUS1	CLUS2	CLUS3	CLUS4	CLUS5
Traditional English	74.55%	98.79%	100.0%	0.00%	0.00%	99.39%
Alternative English	15.41%	0.00%	0.00%	80.52%	38.71%	0.00%
Media Studies	8.24%	11.52%	3.30%	2.60%	6.45%	11.04%
ESOL	5.38%	1.21%	0.00%	5.19%	33.87%	1.84%
Traditional Mathematics	48.21%	2.42%	83.52%	6.49%	61.29%	89.57%
Alternative Mathematics	23.66%	45.45%	6.59%	61.04%	6.45%	0.00%
Accounting	8.24%	4.85%	8.79%	2.60%	17.74%	10.43%
Agriculture/Horticulture	6.27%	9.70%	2.20%	11.69%	1.61%	4.29%
Biology	28.85%	23.03%	91.21%	6.49%	19.35%	14.11%
Chemistry	24.37%	4.24%	70.33%	0.00%	29.03%	28.83%
Electronics	4.66%	4.85%	1.10%	5.19%	8.06%	4.91%
Physics	23.84%	5.45%	34.07%	3.90%	40.32%	39.88%
Physical Education	27.06%	30.91%	27.47%	42.86%	6.45%	23.31%
Health & Lifeskills	7.53%	8.48%	3.30%	7.79%	14.52%	6.13%
Sports	21.86%	29.70%	7.69%	42.86%	16.13%	14.11%
Geography	10.57%	16.36%	14.29%	6.49%	1.61%	7.98%
History	15.41%	20.00%	23.08%	2.60%	3.23%	17.18%
Economics	9.68%	6.67%	4.40%	0.00%	19.35%	16.56%
Tourism & Hospitality	8.42%	12.12%	3.30%	18.18%	1.61%	5.52%
Classics/Latin	9.86%	13.33%	4.40%	0.00%	3.23%	16.56%
European Languages	5.73%	3.64%	9.89%	1.30%	0.00%	9.82%
Te Reo Mäori	4.48%	6.67%	1.10%	9.09%	3.23%	2.45%
Practical Technology	7.53%	6.06%	2.20%	16.88%	4.84%	8.59%
Graphics & Design	13.08%	10.30%	5.49%	9.09%	9.68%	23.31%
Information Management	14.16%	18.18%	8.79%	11.69%	4.84%	17.79%
Computer Studies	18.46%	.73%	20.88%	22.08%	33.87%	15.34%
Music	10.93%	11.52%	.09%	14.29%	9.68%	8.59%
Drama	10.39%	20.00%	2.20%	7.79%	3.23%	9.20%
Visual Arts	13.44%	15.15%	.09%	7.79%	9.68%	16.56%
Photography	8.78%	19.39%	4.40%	0.00%	4.84%	6.13%
Transition	9.14%	12.12%	0.00%	24.68%	9.68%	3.68%
Vocational	27.24%	47.88%	2.20%	38.96%	6.45%	22.70%

The dendrogram (Figure 5) indicates that five clusters will be useful.

Figure 5 Dendrogram for Year 12 clusters



Subjects that characterise Year 12 clusters are set out in Table 7 below.

Table 7 **Year 12 cluster characteristics**

Cluster1 *n = 165*	Cluster2 *n = 91*	Cluster3 *n = 77*	Cluster4 *n = 62*	Cluster5 *n = 163*
Traditional English	Traditional English	Alternative English	Alternative English	Traditional English
Media Studies	Traditional Mathematics	Alternative Mathematics	ESOL	Media Studies
Alternative Mathematics	Biology	Agriculture/Horticulture	Traditional Mathematics	Traditional Mathematics
Agriculture/Horticulture	Chemistry	Electronics	Accounting	Accounting
Physical Education	Physics	Physical Education	Chemistry	Chemistry
Health & Lifeskills	Geography	Sports	Electronics	Physics
Sports	History	Tourism & Hospitality	Physics	History
Geography	European Languages	Te Reo Mäori	Health & Lifeskills	Economics
History	Computer Studies	Practical Technology	Economics	Classics/Latin
Tourism & Hospitality	Music	Computer Studies	Computer Studies	European Languages
Classics/Latin		Music		Practical Technology
Te Reo Mäori		Transition		Graphics & Design
Information Management		Vocational		Information Management
Drama				Visual Arts
Visual Arts				
Photography				
Transition				
Vocational				
Overall descriptions of subjects which characterise the clusters				
Traditional English Alternative mathematics Other practical Arts subjects	Traditional English & mathematics All (3) traditional sciences Other academic	Alternative English & mathematics Alternative science Electronics Other practical	Alternative English Traditional mathematics ESOL Science/ Accounting/IT	Traditional English & mathematics Accounting Science (not biology) Other academic Other practical

In the Year 12 subject choices we see again that clusters are associated with school, ethnic group, and gender. As with the Year 11 data, it is difficult to tell whether these effects are impinging on one another, or whether they are separate effects. What we do know is that school and ethnicity are associated. This goes along with geographically clustered populations and school zoning, so is not unexpected, but does make our results more difficult to interpret. Also, as we have two single-sex schools, school and gender have a significant association, which makes it difficult to isolate gender vs. school effects in subject choice.

The following table (Table 8) shows a clear school effect. Whether the school effect is a school effect *per se* or an obscured gender/ethnic effect is hard to tell. Cluster 1 has a much higher than expected proportion of students from City School B; Cluster 2 has higher than expected proportions of students from City School A and Town School D; Cluster 3 has higher than expected proportions of students from Schools C, E, and F; Cluster 4 has more students than expected from Schools C and F; and Cluster 5 has more students than expected from School A. Each school is strongly represented in just one or two clusters.

Table 8 **School by Year 12 cluster**

School ↓		**CLUSTER 1**	**2**	**3**	**4**	**5**	**Overall**
A	*n*	15	**28**	1	7	**38**	89
	%	9.09	**30.77**	1.3	11.29	**23.31**	15.95
B	*n*	**79**	24	11	16	54	184
	%	**47.88**	26.37	14.29	25.81	33.13	32.97
C	*n*	17	2	**18**	**13**	14	64
	%	10.30	2.20	**23.38**	**20.97**	8.59	11.47
D	*n*	16	**13**	9	5	16	59
	%	9.70	**14.29**	11.69	8.06	9.82	10.57
E	*n*	28	10	**19**	8	29	94
	%	16.97	10.99	**24.68**	12.90	17.79	16.85
F	*n*	10	14	**19**	**13**	12	68
	%	6.06	15.38	**24.68**	**20.97**	7.36	15.40
Total	*n*	165	91	77	62	163	558
	%	29.57	16.31	13.80	11.11	29.21	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	108.08
Df	20
p-value	<.0001
Sample size	558

Table 9 shows that Päkehä students are found mostly in Clusters 1, 2, and 5. Cluster 3, characterised by alternative English and mathematics courses along with more practical subjects, contains a predominance of Pacific and Mäori students. Cluster 4, characterised by alternative English, traditional mathematics, with science, accounting, and IT, is notably populated with Asian students.

Table 9 **Ethnic group by Year 12 cluster**

Ethnic Group ↓		CLUSTER 1	2	3	4	5	Overall
Asian	*n*	5	9	3	**23**	9	49
	%	3.14	10.84	4.11	**37.70**	5.96	9.30
Mäori	*n*	22	2	**23**	7	15	69
	%	13.84	2.41	**31.51**	11.46	9.93	13.09
Pacific	*n*	15	3	**15**	6	9	48
	%	9.43	3.61	**20.55**	9.84	5.96	9.11
Päkehä	*n*	**117**	**69**	32	25	**118**	361
	%	**73.58**	**83.13**	43.84	40.98	**78.15**	70.07
Total	*n*	159	83	73	61	151	527
	%	30.17	15.75	13.85	11.57	28.65	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	124.44
Df	12
p-value	<.0001
Sample size	527

There are more than expected numbers of male students in Clusters 3 and 4 (Table 10), and more than expected numbers of female students in Clusters 1 and 2. Cluster 5 is represented by male and female students in approximately the same proportions as the whole group together. As with the Year 11 data, this effect is probably related to the school effect.

Table 10 **Gender by Year 12 cluster**

Gender ↓		CLUSTER 1	2	3	4	5	Overall
Male	*n*	72	38	**52**	**38**	84	284
	%	43.64	41.76	**67.53**	**61.29**	51.53	50.90
Female	*n*	**93**	**53**	25	24	79	274
	%	**56.36**	**58.24**	32.47	38.71	48.47	48.90
Total	*n*	165	91	77	62	163	558
	%	29.57	16.31	13.80	11.11	29.21	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	17.75
Df	4
p-value	0.0014
Sample size	558

4.3 Year 13 Results

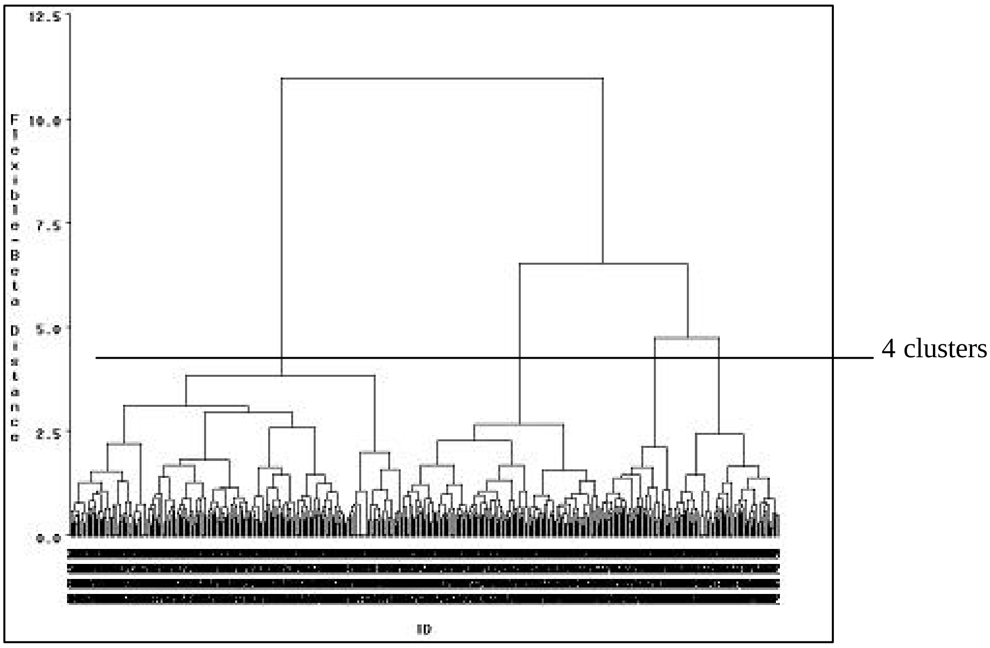
The same process was applied to the Year 13 data. Table 11 shows the overall proportions of cluster membership by subject.

Table 11 **Year 13 clusters**

	Overall	CLUS1	CLUS2	CLUS3	CLUS4
Traditional English	49.08%	0.00%	100.0%	0.00%	100.0%
Contextually-focused English	4.29%	14.00%	0.00%	0.00%	0.00%
Media Studies	10.74%	15.00%	8.75%	10.61%	7.50%
ESOL	6.13%	4.00%	0.00%	24.24%	0.00%
Accounting	7.06%	4.00%	16.25%	4.55%	3.75%
Calculus	27.30%	3.00%	43.75%	69.70%	6.25%
Statistics	35.89%	3.00%	65.00%	75.76%	15.00%
Agriculture/Horticulture	5.21%	8.00%	2.50%	4.55%	5.00%
Biology	27.61%	11.00%	43.75%	40.91%	21.25%
Chemistry	21.78%	1.00%	42.50%	50.00%	3.75%
Physics	25.15%	5.00%	48.75%	54.55%	2.50%
Physical Education	17.48%	19.00%	7.50%	15.15%	27.50%
Sports	9.51%	23.00%	0.00%	4.55%	6.25%
Geography	14.42%	16.00%	12.50%	7.58%	20.00%
History	14.72%	9.00%	16.25%	9.09%	25.00%
Economics	15.34%	6.00%	23.75%	19.70%	15.00%
Tourism & Hospitality	7.67%	18.00%	2.50%	0.00%	6.25%
Classics/Latin	11.04%	6.00%	11.25%	4.55%	22.50%
Graphics & Design	11.35%	9.00%	13.75%	6.06%	16.25%
Information Management	7.67%	13.00%	5.00%	3.03%	7.50%
Computer Studies	8.59%	13.00%	6.25%	12.12%	2.50%
Music	5.52%	7.00%	2.50%	3.03%	8.75%
Drama	7.67%	7.00%	2.50%	4.55%	16.25%
Visual Arts	19.63%	20.00%	13.75%	15.15%	28.75%
Photography	6.75%	4.00%	5.00%	1.52%	16.25%
Art History	4.60%	5.00%	0.00%	0.00%	12.50%
Transition	6.13%	17.00%	0.00%	1.52%	2.50%
Correspondence Subject	5.52%	5.00%	1.25%	3.03%	12.50%
Vocational	16.26%	29.00%	2.50%	3.03%	25.00%

Perhaps as a result of students taking fewer subjects at Year 13, and also that students will be focusing on prerequisites for their chosen futures, the subject data suggest just four clusters at Year 13 (Figure 6).

Figure 6 Dendrogram for Year 13 clusters



Year 13 cluster characteristics are set out below in Table 12. At Year 13 (perhaps because students are focusing their genuine choices more), there seems to be a clearer delineation between cluster characteristics than at other year levels. Cluster 2 includes students who are orientated towards the more practical subjects; Cluster 3 is characterised strongly by the more academic subjects; Cluster 4 is distinctive for its science and computing bias, and we note that students in this cluster are also predominantly ESOL students; Cluster 1 appears to be characterised by more academic arts subjects as well as some practical subjects.

Table 12 **Year 13 cluster characteristics**

Cluster1 *n = 80*	**Cluster2** *n = 100*	**Cluster3** *n = 80*	**Cluster4** *n = 66*
Traditional English	ESOL	Traditional English	Contextually-focused English
Physical Education	Calculus	Accounting	Media Studies
Geography	Statistics	Calculus	Agriculture/Horticulture
History	Biology	Statistics	Sports
Classics/Latin	Chemistry	Biology	Geography
Graphics & Design	Physics	Chemistry	Tourism & Hospitality
Music	Economics	Physics	Information Management
Drama	Computer Studies	History	Computer Studies
Visual Arts		Economics	Music
Photography		Graphics & Design	Transition
Art History			Vocational
Correspondence Subject			
Vocational			
Overall descriptions of subjects which characterise the clusters			
Traditional English	ESOL	Traditional English	Alternative English
Arts subjects	2 mathematics	2 mathematics	Other practical
Other practical	3 sciences	3 sciences	

In Table 13 Cluster 1 is dominated by City School A, Cluster 2 by Schools C and F, Cluster 3 by Schools A and E, and Cluster 4 by City School B. Town School D appears to lie as expected across the clusters. Although the significant p-value for Table 13 indicates an association between school and groups of subject choices, we should check whether this is indeed an isolated school effect, or whether ethnic group or gender effects may be related to this school effect.

Table 13 **School by Year 13 cluster**

School ↓		CLUSTER 1	2	3	4	Overall
A	*n*	**26**	15	**31**	9	81
	%	**32.50**	15.00	**38.75**	13.64	24.85
B	*n*	29	28	12	**29**	98
	%	36.25	28.00	15.00	**43.94**	30.06
C	*n*	5	**16**	2	4	27
	%	6.25	**16.00**	2.50	6.06	8.28
D	*n*	9	10	11	5	35
	%	11.25	10.00	13.75	7.58	10.74
E	*n*	5	14	**21**	14	54
	%	6.25	14.00	**26.25**	21.21	16.56
F	*n*	6	**17**	3	5	31
	%	7.50	**17.00**	3.75	7.58	9.51
Total	*n*	80	100	80	66	326
	%	24.54	30.67	24.54	20.25	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	59.57
Df	15
p-value	<.0001
Sample size	326

In Table 14 we observe again a clear division with respect to ethnicity. However, despite the clear effect, we cannot assume that it is in any way causal. Ethnic populations in New Zealand tend to be geographically clustered, so ethnic proportions within schools are unlikely to reflect national proportions. Whether the effect we observe here is down to choices made by students or due to their ethnic groups or due to different school policies, or both (or neither), is impossible to tell. It is, however, interesting to note how Asian students appear to be leaning towards the sciences (without English), Päkehä students towards either sciences (with English), or the more academic arts subjects, and Mäori and Pacific students towards the practical options available.

Table 14 **Ethnic group by Year 13 cluster**

Ethnic Group ↓		CLUSTER 1	2	3	4	Overall
Asian	*n*	8	4	11	**29**	52
	%	10.67	4.49	15.28	**49.15**	17.63
Mäori	*n*	8	**13**	3	4	28
	%	10.67	**14.61**	4.17	6.78	9.49
Pacific	*n*	4	**13**	4	1	22
	%	5.33	**14.61**	5.56	1.69	7.46
Päkehä	*n*	**55**	59	**54**	25	193
	%	**73.33**	66.29	**75.00**	42.37	65.42
Total	*n*	75	89	72	59	295
	%	25.42	30.17	24.41	20.00	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	65.49
Df	9
p-value	<.0001
Sample size	295

In the following table (Table 15) note that Cluster 1, which is characterised by students taking arts subjects (as opposed to science), is strongly populated by female students. Cluster 4, the cluster that includes many of the Asian students who have chosen science subjects, is somewhat male dominated.

Table 15 **Gender by Year 13 cluster**

Ethnic Group ↓		CLUSTER 1	2	3	4	Overall
Male	*n*	25	48	34	**39**	146
	%	31.25	48.00	42.50	**59.09**	44.79
Female	*n*	**55**	52	46	27	180
	%	**68.75**	52.00	57.50	40.91	55.21
Total	*n*	80	100	80	66	326
	%	24.54	30.67	24.54	20.25	100.00

Note: Bold print shows dominant cluster membership.

Statistics

Chi-square statistic	12.00
Df	3
p-value	0.0075
Sample size	326

4.4 Tangled effects

Now we return to the question of whether we can actually isolate school, ethnic, and gender effects from each other. An initial ploy is to run a Cochran-Mantel-Haenszel test for association. This test gives a stratified statistical analysis of the relationship between two variables after controlling for others, and thus provides a way to adjust for possible confounding effects. For example, we may wish to know whether there is a relationship between cluster and ethnic group after controlling for school and gender. SAS (SAS Institute Inc, 1999–2001) provides the statistic we need in the form of a "general association statistic" generated by the `freq` procedure. This statistic is used where variables are nominal. We test the null hypothesis of no association between cluster and ethnic group in any stratum, against the alternative hypothesis that for at least one stratum there is some kind of relationship.

The results for Year 11 are shown in Table 16. There is no gender effect after allowing for ethnic group and school. This implies that the gender effect observed in the results in Section 3.1 is tied up in school and ethnic group effects (probably mostly school). In other words, we do not have evidence of a subject choice gender effect at Year 11. Probably the gender effect observed earlier is due, at least in part, to the two single-sex schools being in the sample of schools.

There is, however, a school effect after allowing for ethnic group and gender. That is, the data support the hypothesis that in at least one ethnic-by-gender stratum we are seeing an association between cluster and school. The significant χ^2-statistic for ethnic group indicates that in at least one school-by-gender stratum there is a relationship between cluster and ethnic group.

Table 16 **Cochran-Mantel-Haenszel test results for Year 11**

Testing for association between cluster and…	Controlling for	df	c^2-statistic	Prob c^2
school	ethnic group gender	35	205.77	<.0001
ethnic group	school gender	21	64.86	<.0001
gender	school ethnic group	7	6.14	0.5234

Effective sample size = 586
Frequency missing = 29

Table 17 gives the results for the Year 12 data. There does appear to be a gender effect at Year 12, after allowing for school and ethnic group. Whether this effect reflects reality or whether the effect is generated by sample idiosyncrasies is difficult to tell. In general, we can be more confident about effects which are shown consistently across year levels, which the gender effect is not.

Table 17 **Cochran-Mantel-Haenszel test results for Year 12**

Testing for association between cluster and…	Controlling for	df	c^2-statistic	Prob c^2
school	ethnic group gender	20	118.27	<.0001
ethnic group	school gender	12	147.86	<.0001
gender	school ethnic group	4	21.81	0.0002

Effective sample size = 527
Frequency missing = 20

Table 18 gives the results for Year 13. These results are similar to those for the Year 11 cohort. There is no gender effect to be seen after allowing for school and ethnic group. This lends weight to the point made above that the gender effect seen at Year 12 may indeed be a sample anomaly.

Table 18 **Cochran-Mantel-Haenszel test results for Year 13**

Testing for association between cluster and…	Controlling for	df	χ^2-statistic	Prob χ^2
school	ethnic group gender	15	45.00	<.0001
ethnic group	school gender	9	62.39	<.0001
gender	school ethnic group	3	0.92	0.8205

Effective sample size = 295
Frequency missing = 25

If we had enough observations in our sample we might pursue log-linear models to model the frequencies in the 4-way table cluster * school * ethnic group * gender. These models would allow us to explore the interactions which undoubtedly exist between the explanatory variables. However, a standard rule of thumb for log-linear models is that one must have at least five observations for every cell. For the Year 11 data that amounts to a minimum of 5(observations) x 8(clusters) x 6(schools) x 4(ethnic groups) x 2(genders) = 1920 observations in all – which we certainly do not have! For this reason we have not investigated the log-linear models in this part of the Learning Curves project.

5. Conclusion

We clustered students within year level according to their subject choices. Students with similar choices of subjects were grouped together by a standard clustering procedure. Clusters are characterised by certain subjects (those taken by students in the cluster). That is, students in a particular cluster have a comparatively high probability of taking subjects characterising that cluster.

The question we aimed to answer was: Are these groups (clusters) associated with other (demographic) variables in the dataset? Answering this may help to throw some light on answers to further naturally arising questions: Do different schools have different policies regarding subject selection for their students, or different expectations or biases which affect student subject choice? Do different cultures have different expectations or perceptions which take effect in the home, and at school with respect to subject choice? And, of course, the age-old question of whether subject choice is gender specific.

While we cannot actually *predict* cluster membership from the demographics in this dataset, there are some interesting patterns to be observed. It is important to note that the observed patterns cannot lead us to any generalised conclusions about the nature of the relationship between subject choice and the demographic variables. First, we do not have a sample representative of a wider population, so we cannot make inferences about, for example, what is happening on a national or even regional level. Second, due to the nature of the sample, we are not able to effectively extend the research to include log-linear models from which we might extract information about the interactions between school, ethnic group, and gender. However, that clear patterns of subject choice merely exist (the clustering procedure produced well-defined clusters) is interesting, and further, that the identified clusters bear strong relationships to all the demographic variables available is also a matter of great interest, and points to possibilities for further research into the nature of the associations between the subjects students choose to take at school and their demographic profiles.

In the current environment of increasing ability to store, retrieve, and share information at a school, regional, and national level, we could perhaps begin to make use of available administrative data to answer some pertinent questions about subject choices (or groups of subject choices) with respect to differences in school policies and perceptions, expectations which (rightly or wrongly) relate to cultural background, gender differences and/or biases, and the effects of various socioeconomic factors.

References

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis*. London: Arnold.

Hipkins, R., Vaughan, K., with Beals, F., Ferral, H., & Gardiner, B. (2005). *Shaping our futures: Meeting secondary students' needs in a time of evolving qualifications*. Wellington: New Zealand Council for Educational Research.

Kaufmann, L., & Rousseeuw, P. J. (1990). *Finding groups in data*. New York: John Wiley & Sons, Inc.

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Computer Journal*, *9*, 373–380.

Sneath, P. H. A. (1957). Some thoughts on bacterial classification. *Journal of General Microbiology*, *17*, 184–200.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*, 236–244.

Statistical programs

SAS Institute Inc. (1999–2001). Version 8.02 of the SAS System for Windows. Cary, NC, USA: SAS Institute Inc.