The Classification of Students in Two Multiple Choice Question Examinations by Numerical Taxonomy

© 1993, M.D.Buckley-Sharp
Department of Chemical Pathology
University College London Hospitals

Editorial Note: This paper was originally prepared for publication in 1993 after a lengthy gestation. It was submitted to and refereed by Medical Education but declined as it could not stand alone without its companion paper (see comments there). Some of the referees' minor comments have been adopted when preparing this transcription.

As far as could be determined then, and I have seen nothing since to change my view, this remains an entirely unique perception on a use of the data from Multiple Choice Question examinations.

(2005) M.D.Buckley-Sharp

Summary

This paper uses the data from two multiple choice question examinations in Medicine. The raw response data is analysed using standard methods of numerical taxonomy. Various solutions are proposed to the classification of the participants, who are the students as well as the examiner. The solutions demonstrate stability of the classification between the two examinations: interpretation is improved by recognising a null group who had sat the second examination, but not the first. Some features of the proposed groups are described, including the possibility that some differences are due to secondary skills in taking examinations in the multiple choice question format.

Key words:
Numerical Taxonomy
Cluster Analysis
Systematics
Multiple Choice Questions (MCQ)

Introduction

The standard procedure in multiple choice question (MCQ) examinations, as in all other types of examination, is to derive a score for each student or candidate. By whatever method, this is a unidimensional score. It is assumed that two students with the same score have performed identically, and that where two students have different scores then one is better than the other.

Buckley-Sharp (1993) proposed that unidimensional scoring is a limited view of the data, and that the methods of numerical taxonomy have the potential to elucidate groups or clusters of participants. Such a classification might have value for diagnostic, therapeutic, or prognostic purposes in education.

Earlier work by Buckley-Sharp et al (1969) used available scoring parameters as continuous variables for clustering. In a simplistic way, the setting of examinations simultaneously in different subjects, whether at secondary or tertiary educational level, is also a recognition that students may not perform equally well at everything.

Student performances are not homogeneous, but neither are they completely heterogeneous ie, absolutely individual. More likely, a limited number of performance patterns exist within any group of students.

It is suggested that the raw response data from MCQ examinations can be used directly for numerical taxonomy. The raw data is binary, and of high dimensionality. Although the calculations in numerical taxonomy are not complex, the quantity of calculation is high and of order $m^2.n$ (where m is the number of participants, and n is the number of data items). The feasibility of computing a taxonomy is more limited by the number of participants, while the large number of data items is advantageous and tolerable.

For a classification of students to have any value, it must first be shown to have some stability. Stability is shown by a repetition of the same classification memberships from a separate data collection. Without prior experience, stability would seem more likely if two collections of data are close in time, and are in the same subject area. If there were no classification stability even with these restrictions, then stability is unlikely for other more relaxed comparisons. For example, it would seem quite unlikely that school Physics and final Medicine should each demonstrate the same taxonomy of students, although this has never been tested.

This paper has two purposes. It provides examples of the methods of numerical taxonomy of MCQ examinations data: expanding on the review by Buckley-Sharp (1993). It also seeks to demonstrate stability of classification results, albeit under tight restrictions.

<<Figure 1 of the original paper is omitted, and references to it have been removed>>

Data

The first MCQ examination studied was a pre-final test in Medicine (Pretest). There were 60 questions in the five-alternative true-false format, making 300 alternatives in all. Responses were collected by the two-box method ie, one box for 'select if true' and an associated box for 'select if false'. Therefore, there were 600 response boxes.

The second MCQ examination was a final-MB (University of London) test in Medicine (Final). There were 100 questions in the same format and with response data collected by the same method as the Pretest. Therefore there were 1000 response boxes.

These two examinations were taken only a few weeks apart. The concentration by subject and time was chosen deliberately to see whether cluster stability could exist.

Candidates taking both examinations were matched by name, and the sex of each candidate was recorded. Some candidates at this school were preparing for the Oxford BM degree and took only the Pretest. Some candidates for the London MB degree did not take the Pretest. Therefore each examination had a null group of candidates: those who did not take the examination, but did take the other.

Methods

Data collection was via mark-sense cards, and both examinations were scored conventionally using the SCORE9 program (Buckley-Sharp & Harris (1971)). Standard examination and question analysis statistics were obtained (Buckley-Sharp & Harris (1972)).

SCORE9 is part of a much larger suite (called UT3), which allows examination response data collected in various ways to be translated to a common format, and held in a library for further analysis. UT3 already contained a wide and necessary range of support routines. It was sufficient to add three more respectively,

- To create a matrix of distance coefficients;
- To run a clustering algorithm; and
- To display a dendrogram used to illustrate the clustering results.

The numerical methods used were those described by Buckley-Sharp (1993). The objects in the analysis were all the students/candidates and including the examiner. Coefficients between each pair of objects were calculated as the average squared Euclidean distance. This distance is simply the proportion of binary disagreements between the pair of objects, and is the same as the Hamming distance.

Cluster analysis was performed using Ward's Incremental Sum of Squares (ISS) method. At each cluster cycle, one pair of objects or intermediate clusters was chosen for fusion, where the corresponding distance coefficient was the minimum in the current matrix. The distance coefficients from the new cluster to all other remaining objects/clusters were then recalculated, and the matrix reduced. Cycles by this method are iterative. The m objects start, one each, in m clusters. After a total of (m-1) fusion steps, all objects are finally in the same cluster.

The identities of the chosen cluster fusion-pair and their pre-fusion distance coefficient were recorded at each fusion step. The resulting list, by fusion step number, was used to organise the display of a dendrogram of the clustering process. The fusion distance function was used as the axis of the dendrogram.

Each student was identified by a candidate number, with the examiner as candidate 'zero'. Each fusion was described using these numbers, with the successor fused cluster adopting the lower candidate number. The examiner must therefore retain the 'zero' label throughout, and the last fusion cycle must merge to the examiner. The examiner appears along one complete edge of the display dendrogram. While this is a convention, and not meant to give unique properties to the examiner, it does have the merit that the examiner is separately regarded as having the perfect score. Therefore, students who cluster early to the examiner, and hence are shown near to the examiner's edge on the dendrogram, are likely to be those with the highest scores. The presentation of results which follows uses letter labels for the last few clusters. It is still the case that the examiner is in the cluster with the first label.

The recommendation of a cluster solution remains arbitrary. It is likely to be within the last few cluster fusion cycles ie, having relatively few clusters. Youngman (1979) is merely one of many to use a rapid rise in the fusion distance function as a criterion.

In this study, which is searching also for stability between taxonomies obtained from two sources, use has been made of Kappa, the coefficient of agreement, proposed by Cohen (1960), and cited by Youngman (1979). Two cluster solutions for comparison are set out in a c x c table, as in Table 1, where c is the number of clusters in both solutions. Each cell contains the count of original objects in each cluster of the two solutions: the rows and columns are shuffled to maximise the diagonal elements.

Table 1: An example of a 3x3 cluster comparison contingency table, with marginal totals. Values in the table are counts of matched objects in each cluster of two proposed solutions, labeled (L,M,N) and (X,Y,Z).

Solution-1 clusters	**L**	**M**	**N**	
Solution-2 clusters				
X	25	5	7	37
Y	6	24	8	38
Z	4	9	23	36
	35	38	38	111

Then, $K = (P_o - P_c) / (1 - P_c)$

where K is the Kappa coefficient of agreement;

P_o is the observed on-diagonal agreement ie, the sum of the on-diagonal elements divided by the total objects;

P_c is a correction for chance agreement which is found by the standard chi-square method. For each on-diagonal element the chance expected value is

<row-total> x <column-total> / <table-total>

and the chance agreement is then the sum of these on-diagonal estimates divided by the total objects

And, for the values in Table 1,

$P_o = 72 / 111 = 0.649$

$P_c = \{ (37 \times 35)/111 + (38 \times 38)/111 + (36 \times 38)/111 \} / 111 = 0.333$

$K = (0.649 - 0.333) / (1 - 0.333) = 0.47$

Kappa ranges from 0 to 1. There is no formal test for statistical significance ie, improvement over chance: this work is looking instead for good agreement. Vogt et al (1987) includes a further discussion of Kappa, and also of extensions to several other methods for comparing cluster solutions.

Notice that in the calculation of P_c, all off-diagonal elements (disagreements) have equal status. Although information may be available on the spatial distribution and distances between particular clusters, no use is made of this in calculating P_c, and hence Kappa. Shuffling the rows and columns (Table 1) to maximise the on-diagonal elements will maximise Kappa. However, the analyst might be faced with several possible arrangements. A sub-maximal Kappa might be preferred if a particular match of solutions was considered more relevant after reviewing the actual relationships of clusters. (Referees note: there is a version of Kappa allowing off-diagonal weighting.)

There is a restriction, when using Kappa, to pairs of proposed solutions with equal numbers of clusters. However, it will be seen later how a null cluster can help in the analysis.

Results

The basic scoring results from the two examinations are shown in Table 2. Both examinations had high reliability, presumably because they were quite long.

Table 2: Basic results from the two examinations studied.

Examination		Pretest	Final
Questions	MCQ	60	100
	Response boxes	600	1000
Students	Both exams	73	73
	Pretest only	4	
	Final only		13
	Total	77	86
Scores. mean(sd)	%Correct	62.2 (11.2)	66.3 (10.0)
	%Error	14.1 (5.4)	(12.4 (4.6)
	%Adjusted	48.1 (12.5)	53.8 (10.4)
Reliability	KR20	0.94	0.95
Cluster objects	Examiner	1	1
	Students	77	86
	Total	78	87
	Both exams	74	74

There were 74 objects common to both examinations. Four additional objects in the Pretest are sufficiently few in number to be regarded as passengers in the process and unlikely to distort the clustering process. However, it will be seen later how interpretation of the cluster results was aided by considering the 13 additional objects in the Final examination as a separate null group in the Pretest.

Figure 2: Graphs of the fusion distance functions for the last twenty fusion steps. Pretest and Final examinations.

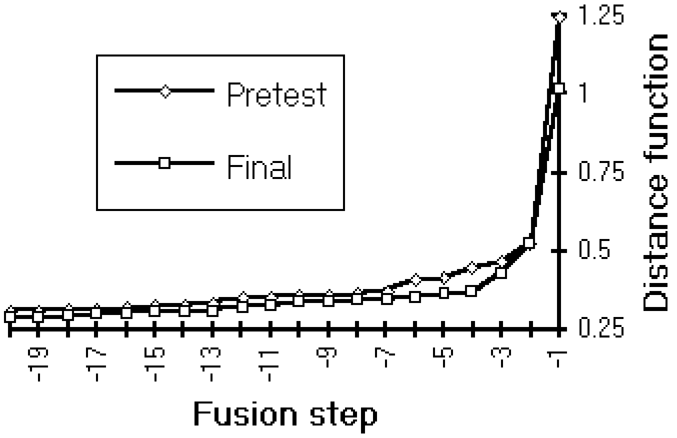

Figure 3: Dendrogram showing fusions from the eight-cluster state in the Pretest. The numbers for each cluster **A** to **H** show the membership size: at each fusion, memberships are added. Numbers in brackets show the subset of members who also sat the Final examination.

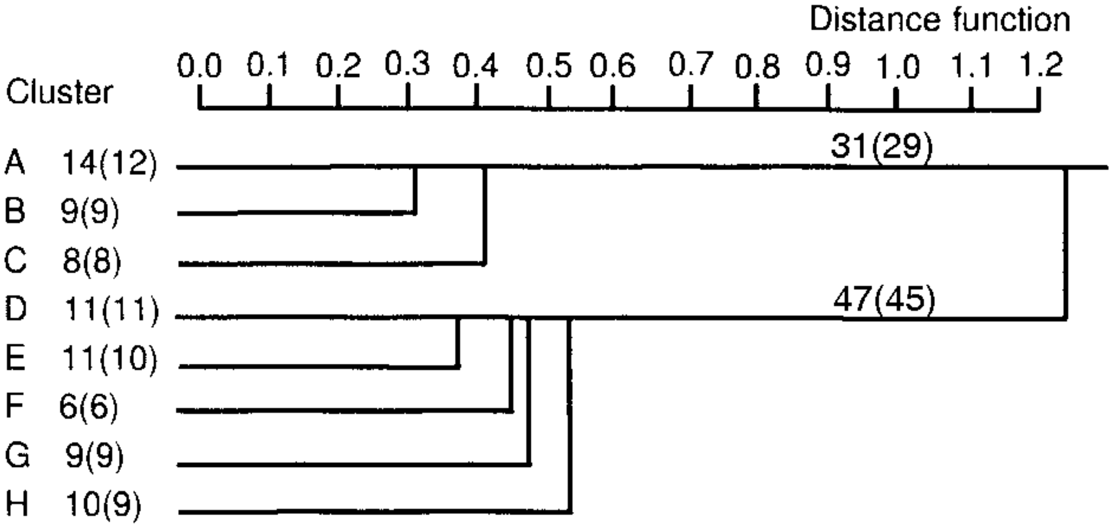

Figure 4: Dendrogram showing fusions from the eight-cluster state in the Final. The numbers for each cluster **O** to **U** show the membership size: at each fusion, memberships are added. Numbers in brackets show the subset of members who also sat the Pretest examination.

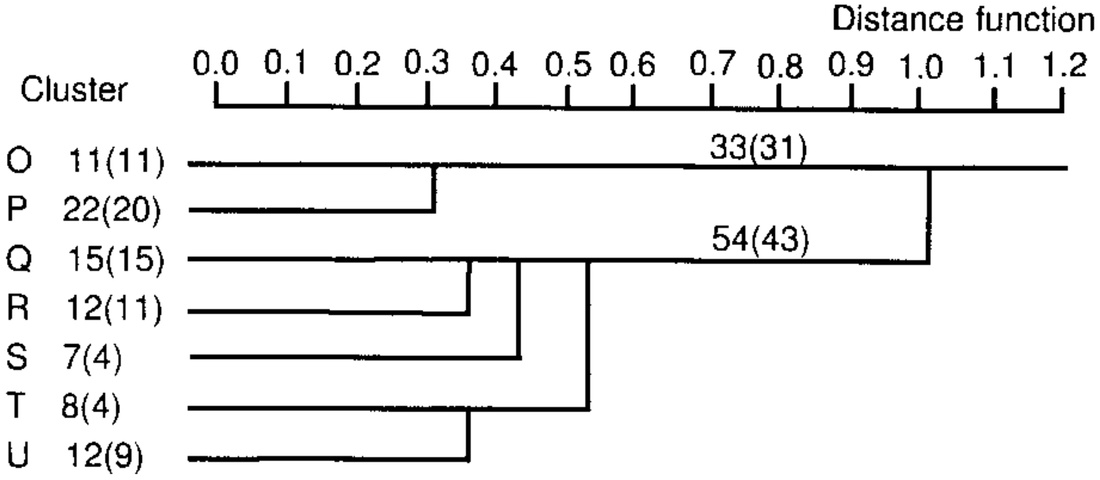

Figure 2 shows the successive values of the fusion distance functions. The two examinations had different numbers of objects, and therefore different numbers of fusions steps, but the last twenty steps are shown in each case. Both fusion functions show the expected monotonic rise. This is most marked for the last fusion cycle when all objects join together into one cluster, and suggests that at least two clusters exist in each examination. Whether there are more clusters, up to about 4 in the Final examination, is more debatable.

Figure 3 shows the dendrogram starting at an eight-cluster state for the Pretest. Pretest clusters have been labeled **A** to **H**. Figure 4 shows the dendrogram starting at a seven-cluster state for the Final. Final clusters have been labeled **O** to **U**. These partial solutions were chosen as starting points for further study. Both have five-membered lower cluster groups. The upper groups differ, with **C** an apparent orphan cluster.

As noted above, fusion of clusters into the lower labeled object is arbitrary, and Figure 4 would look even more like Figure 3 if clusters **T**,**U** were really matched to **D**,**E** (or **E**,**D**). However, Table 3 suggests that this is not appropriate, since joint cluster membership shows that **D**,**E** are preferentially related to **Q**,**R**.

Table 3: Joint cluster membership for 74 objects participating in both examinations. Clusters labels as in Figure 3 (Pretest) and Figure 4 (Final). Quadrants divide major cluster groups.

Final	**O**	**P**	**Q**	**R**	**S**	**T**	**U**
Pretest							
A	6	5	0	0	0	0	1
B	2	5	0	1	0	0	1
C	3	3	0	0	0	1	1
D	0	2	6	2	0	0	1
E	0	0	3	5	1	0	1
F	0	2	1	1	0	0	2
G	0	3	2	1	1	2	0
H	0	0	3	1	2	1	2

The upper left quadrant in Table 3 contains orphan cluster **C**. To meet the requirement for symmetry, **C** has been joined to **B** when calculating Kappa. The lower right quadrant could be optimised further by transposing **F** and **H**, but this does not fit with later consolidations. The value of Kappa for this table is sub-maximal.

Table 4 shows the comparison statistics for Table 3, and also for subsequent consolidations.

Table 4: Statistics for the various contingency tables shown. **S_d** is the average fusion distance function. **P_o** is the observed proportional agreement. **P_c** is the correction for expected agreement. **K** is Kappa, the coefficient of agreement.

From	Clusters	**S_d**	**P_o**	**P_c**	**K**
Table 3	7x7	0.357	0.392	0.162	0.27
Table 5	4x4	0.410	0.622	0.320	0.44
Table 6	3x3	0.449	0.676	0.383	0.47
Table 7	2x2	0.528	0.838	0.518	0.66
Table 8	3x3	n/a	0.690	0.363	0.51

It would be possible to step down through each possible cluster consolidation, but the next solution which seems worthy of review is the 4x4 arrangement shown in Table 5, for which the statistics are shown in Table 4. Cluster **H** is again ambiguous, because the same value for Kappa would be obtained if **H** was transposed with **G**.

Table 5: Joint cluster membership for 74 objects participating in both examinations. Clusters labels as in Figure 3 (Pretest) and Figure 4 (Final). Consolidation of Table 3 into 4x4 cluster solution.

Final	**O,P**	**Q,R**	**S**	**T,U**
Pretest				
A,B,C	24	1	0	4
D,E,F	4	18	1	4
G	3	3	1	2
H	0	4	2	3

Table 6: Joint cluster membership for 74 objects participating in both examinations. Clusters labels as in Figure 3 (Pretest) and Figure 4 (Final). Consolidation of Table 5 into 3x3 cluster solution.

Final	**O,P**	**Q,R,S**	**T,U**
Pretest			
A,B,C	24	1	4
D,E,F,G	7	23	6
H	0	6	3

Table 7: Joint cluster membership for 74 objects participating in both examinations. Clusters labels as in Figure 3 (Pretest) and Figure 4 (Final). Consolidation of Table 6 into 2x2 cluster solution.

Final	**O,P**	**Q,R,S,T,U**
Pretest		
A,B,C	24	5
D,E,F,G,H	7	38

Table 6 shows a 3x3 arrangement. Here, cluster **H** is left on its own. The 2x2 arrangement is an obvious possible solution, and is shown in Table 7. The high value of the distance function increment (Table 4) suggests that at least two real distinct clusters are present in both examinations.

The 13 students who sat the Final, but did not sit the Pretest, were then reconsidered. Their cluster membership was not evenly distributed amongst the other Final groups. To use them, while allowing for the calculation of Kappa, it was necessary to have only two Pretest clusters, but to use a three-cluster Final. These results are shown in Table 8.

Table 8: Joint cluster membership for 87 objects; 74 participating in both examinations, and 13 taking only the Final (null group in Pretest). Clusters labels as in Figure 3 (Pretest) and Figure 4 (Final).

Final	**O,P**	**Q,R,S**	**T,U**
Pretest			
A,B,C	24	1	4
D,E,F,G,H	7	29	9
null group	2	4	7

To evaluate these possible solutions further, it is necessary to return to external criteria. The sex had been recorded for each student. Table 9 shows that sex is correlated with the two cluster solutions for both Pretest and Final. In this analysis, all students are included, and only the examiner is omitted.

Table 9: Relationship between sex and cluster membership for Pretest and Final examinations. Clusters labels as in Figure 3 (Pretest) and Figure 4 (Final).

Pretest	Male	Female	Final	Male	Female
A,B,C	12	18	**O,P**	14	18
D,E,F,G,H #	33	13	**Q,R,S,T,U**	37	17
Chi-square	7.3			5.2	
p	<0.01			<0.03	

# unknown sex, 1 omitted

Figure 5: Cluster membership (students only) displayed according to adjusted percentage score for both Pretest and Final. **A** refers to clusters **A**,**B**,**C**; **D** refers to clusters **D**,**E**,**F**,**G**,**H** (Figure 3). **O** refers to clusters **O**,**P**; **Q** refers to clusters **Q**,**R**,**S**; **T** refers to clusters **T**,**U** (Figure 4).

PreTest		Final		
A		O		
A		O		
AAA		O		
AA		O		
AAAAA		OO		
AA		O		
A		OO		
AA		O		
AAA	D	OO	Q	
AAA		O		
A	D	OOOO		
	DDDD	OO		
AAA	DD	OO	Q	
	DD	OOOO		
	DD	OOO	Q	
A	DD	OO	Q	T
A	DD		Q	T
A	DDDDD		Q	
	DDD	O	Q	T
	DD	O	Q	TT
	DD		Q	
	DDD			TT
	DD		QQ	TTTT
	DDD		Q	T
	DDD		QQQ	
	D		Q	T
	DD			T
	D		QQ	
	D		QQ	TT
	DD		Q	T
	D		Q	T
			QQQ	
			QQQQ	
				T
				T
			QQ	
			Q	
			Q	
			Q	

Figure 5 shows cluster membership according to the adjusted percentage scores obtained by students. Clusters **A** and **O** contain the examiner for the Pretest and Final respectively, so that these clusters obviously include the highest student scores. There is considerable overlap of scores between clusters, especially between clusters **QRS** and **TU**. Are **QRS** and **TU** really similar, or can they be distinguished?

Table 10: Relationship between score parameters and cluster membership for all Pretest students. Cluster labels as in Figure 3. Scores are shown as mean (s.d.).

Cluster(s)	**A,B,C**	**D,E,F,G,H**	
Members	30	47	
Scores mean(sd)			
%Correct	73.4 (5.5)	55.0 (7.6)	**ABC > DEFGH**
%Error	14.2 (4.2)	14.0 (6.1)	**ABC = DEFGH**
%Adjusted	59.2 (6.9)	41.0 (9.9)	**ABC > DEFGH**
%Decisions	87.6 (6.9)	69.0 (9.9)	**ABC > DEFGH**

Table 11: Relationship between score parameters and cluster membership for all Final students. Cluster labels as in Figure 4. Scores are shown as mean (s.d.).

Cluster(s)	**O,P**	**Q,R,S**	**T,U**	
Members	32	34	20	
Scores mean(sd)				
%Correct	76.4 (4.7)	57.1 (6.2)	65.7 (4.5)	**OP>TU>QRS**
%Error	12.5 (3.3)	10.2 (4.4)	16.5 (4.5)	**OP=QRS<TU**
%Adjusted	63.9 (6.1)	46.9 (8.4)	49.3 (5.5)	**OP>TU=QRS**
%Decisions	88.9 (6.1)	67.2 (8.4)	82.2 (5.5)	**OP>TU>QRS**

Table 10 shows more detailed score parameters for the Pretest clusters (two-cluster solution). Differences were confirmed by Student's t test (not shown). Cluster **ABC** members obtained their higher adjusted percentage scores through making more decisions. A higher proportion of their decisions were correct. Alternatively, it is seen that both groups have very similar error scores.

Table 11 shows more detailed score parameters for the Final clusters (three-cluster solution). Differences were confirmed by Snedecor's F test (not shown). The pattern for cluster **OP** members, including the high percentage decisions, matches that shown previously for cluster **ABC** members. Likewise, the pattern for cluster **QRS** is similar to cluster **DEFGH**.

Then it is seen that cluster **TU** members show a definite third pattern. Their high percentage decisions are like cluster **OP**. Their low scores (adjusted percentage) are like cluster **QRS**. Their discrepancy is seen in their high percentage errors. This pattern is not seen in the Pretest clusters because cluster **TU** seems to equate to those students who did not take the Pretest.

Discussion

The first objective of this paper was merely to demonstrate the use of numerical taxonomy when applied to MCQ data. Buckley-Sharp (1993) had proposed a particular selection from the methods developed generally for numerical taxonomy, and this paper exemplifies their use. Given the availability of mechanised data collection, the methods for numerical taxonomy are not difficult to implement. However, very large data arrays have to be handled.

Numerical taxonomy is renowned for proposing structure even in random data, and caution must be exercised in interpretation. Buckley-Sharp et al (1969) proved the significance of a structure of students in an MCQ examination by performing discriminant analysis on the original data using cluster identity as the criterion variable. The original data in that paper was sectional subscore data: continuous and of few dimensions. Such a process would be more difficult here using binary data of very high dimensionality.

The second objectiove of this paper was to see if there was any commonality between structures found independently in two separate examinations. This might indicate that a valid or relevant structure had been found. Both Buckley-Sharp et al (1969) and Buckley-Sharp (1993) have discussed the desirability of demonstrating structural stability in a group of students. Stability is required for confidence in 'diagnosis', 'management' and 'prognosis'.

The selection of two examinations which were very close in both subject matter and timing was deliberate, and they were chosen to make discovery of a matching structure as likely as possible. Had no adequate agreement been found here, then there would be little point in widening the criteria for any future study. It would have to be doubted whether numerical taxonomy has anything to offer in the evaluation of MCQ data. Given the high reliability of MCQ data, it would then also be unlikely that any other metric could be a better substitute.

Having found a structure it is always tempting to try to apply labels. The following should be view with caution.

Pretest clusters **ABC** and matched Final clusters **OP** are characterised by high decision making, moderate errors, and higher overall scores. With these properties, these clusters might be labeled 'INFORMED'.

Pretest clusters **DEFGH** and matched Final clusters **QRS** are characterised by low decision making, also moderate errors, and so lower overall scores. These clusters might be labeled 'INDECISIVE'.

Those who did not take the Pretest are a reasonable match with Final clusters **TU**. They are characterised by high decision making, but these decisions are often wrong, and so they have lower overall scores. These clusters might be labeled 'INDIFFERENT'.

It is important to emphasise that these labels are not intended to be true in every nuance, or to apply exactly to all individual cluster members. Their value lies more in suggesting patterns of behaviour which might be used to guide the educational process. It may equally be that these behaviours refer particularly to the examination process, or even especially to MCQ.

Unpublished work by Harris (DRSE, see Acknowledgements) has shown that students are less likely to come to a decision in an MCQ examination when faced with a false alternative than with a true alternative. This failure, giving rise to the suggestion of indecisiveness, lowers overall scores. It is important to warn students about this property of MCQ examinations, and to advise them to use as much knowledge as they can muster to come to decisions on as many questions as possible.

The 'Indifferent' group reach plenty of decisions, but have more errors. Could they lack some inbuilt warning which, in the other groups, stops students from making excessive errors? This is not to stray into the extensive discussions which have raged for years into the matter of guessing in MCQ examinations, but perhaps many students do only make decisions up to a point where they become uneasy about the possibility of making errors. In this data, that point is at about an error score of 15%. The 'Indifferent' group presses on regardless, although making little effective headway in their overall score.

Characterising the groups according to their score parameters is not the only possible approach. They could be characterised instead by the subset of questions which discriminate them.

Conventional question analysis is internal to the examination and uses a high/low overall score dichotomy as the criterion (Buckley-Sharp & Harris, 1972). An examination with two clusters allows the usual criterion to be easily replaced. An examination with three clusters gives three pairs, leading to three analyses. With increasing numbers of clusters this rapidly becomes complex, so that parsimonious cluster solutions are desirable here also.

Performing a standard question analysis using cluster membership criteria will distinguish those alternatives which contribute significantly from those which do not. Further information might then be available by considering the subject matter of the discriminating questions. Such an analysis has not been done on the data shown here.

This paper has shown that a taxonomic structure can be observed amongst students taking MCQ examinations. In these examples some possible interpretation of the structure can be offered. Nevertheless this interpretation is limited. If students are 'Indecisive' or 'Indifferent' then these may be general traits worthy of some action. On the other hand, to say that a group is 'Informed' says nothing about how they came to be informed, or even what exactly they are informed about.

Acknowledgements

I thank Professor S.G.Semple, formerly Head of the Department of Medicine, Middlesex Hospital Medical School, for permission to use the examinations data. Also, Mr.F.T.C.Harris and other colleagues in the Department of Research and Service in Education (DRSE), Middlesex Hospital Medical School, for help and encouragement during the lengthy planning and completion of this work.
Note: The cluster analysis software was developed and the data analysed in Dec 1982.

References

Buckley-Sharp M.D., Hamlyn A.N., Harris F.T.C., (1969), The Use of Cluster Analysis to Group Examination Candidates, British Journal of Medical Education, **3**, 225-231.

Buckley-Sharp M.D., Harris F.T.C., (1971), The Scoring of Multiple Choice Questions, British Journal of Medical Education, **5**, 279-288.

Buckley-Sharp M.D., Harris F.T.C., (1972), Methods of Analysis of Multiple Choice Examinations and Questions, British Journal of Medical Education, **6**, 53-60.

Buckley-Sharp M.D., (1993), The Principles of Numerical Taxonomoy applied to Multiple Choice Question Examinations Data, (in preparation).

Cohen J., (1960), A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement, **20**, 37-46.

Examinations Data, Department of Research and Service in Education (DRSE), Middlesex Hospital Medical School.
Pretest. DRSE reference K071045. March 1981
Final. DRSE reference K122020. June 1981

Sneath P.H.A., Sokal R.R., (1973), Numerical Taxonomy, The Principles and Practice of Numerical Classification, Freeman & Co, London.

Sokal R.R., Sneath P.H.A., (1963), Principles of Numerical Taxonomy, Freeman & Co, London.

Vogt W., Nagel D., Sator H., (1987), Cluster Analysis in Clinical Chemistry: A Model, Wiley & Sons, Chichester.

Youngman M.B. (1979), Analysing Social and Educational Research Data, McGraw-Hill, London.